Pseudo-dyadic "interaction" on Amazon's Mechanical Turk

Amy Summerville
Christopher R. Chartier

Miami University

Author's Uncorrected Copy.
Published version available at http://link.springer.com/article/10.3758%2Fs13428-012-0250-9

Author note: Address correspondence to Amy Summerville, Department of Psychology, 90 N. Patterson Ave., Oxford, OH 45056; tel. 513-529-6126; email summera@muohio.edu.

Abstract

Psychological researchers have begun to utilize Amazon's Mechanical Turk (MTurk) marketplace as a participant pool. Although past work has established that MTurk is well-suited to examining individual behavior, pseudo-dyadic interactions, in which participants falsely believe they are interacting with a partner, are a key element of social and cognitive psychology. The ability to conduct such interdependent research on MTurk would increase the utility of this online population to a broad range of psychologists. The current research therefore attempts to qualitatively replicate well-established pseudo-dyadic tasks on MTurk in order to establish the utility of this platform as a tool for researchers. We find that participants do behave as if a partner is real, even when doing so incurs a financial cost, and that they are sensitive to subtle information about the partner in a minimal-groups paradigm, supporting the use of MTurk for pseudo-dyadic research.

Keywords: interaction, internet, cooperation, social influence, interdependent decision making

Pseudo-dyadic "interaction" on Amazon's Mechanical Turk

Amazon's Mechanical Turk (MTurk) website, in which "workers" complete brief online tasks for small payments, has become an enormously popular pool of participants (e.g., Mason & Suri, 2012; Rand, 2011), following papers validating the population and the quality of data it provides for survey research (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler & Ipeirotis, 2010). MTurk addresses several key needs of researchers: a large, relatively diverse sample (compared to most undergraduate participant pools) that can provide data quickly and at low cost. Although MTurk is well-suited for survey and experimental research focused on individuals, as noted above, expanding the range of studies that can appropriately be run on MTurk would be quite useful to the field. In particular, many psychologists wish to investigate behavior in an interactive or social context, such as judgment and decision-making (JDM) researchers interested in interdependent decision-making or social psychologists interested in interpersonal interactions. Such studies often require participants to believe that another person is currently interacting with them (e.g., Pillutla & Murninghan, 1996; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Williams, Cheung, & Choi, 2000). This research therefore aims to examine whether research involving ostensibly dyadic tasks can be conducted on MTurk with confidence. We wanted to determine whether participants behave as if their "partners" are real by establishing that the types of effects previously found in lab research will also occur in online studies with MTurk participants.

We focus specifically in the current studies on pseudo-dyadic tasks, research that only purports to involve interaction, rather than attempting to actually allow direct interaction between participants. Research involving only a simulacrum of interaction is quite common throughout psychology. In many studies, although participants may be in the lab simultaneously

with another person (either a confederate or another participant), the content of the ostensible interaction is in fact predetermined by the experimental protocol. Paradigms that involve ostensible interaction are widely used in the social rejection literature (see Blackhart et al., 2009 for a meta-analysis). One commonly used paradigm of social rejection, Cyberball (Williams, Cheung, & Choi, 2000), involves participants playing a computerized game of "catch" that is ostensibly with other participants but in fact is entirely pre-programmed. Similarly, research on economic decision-making frequently asks participants to make and respond to offers from phantom partners. For example, participants may be asked to respond to bargaining offers that are ostensibly made by a partner, but are instead predetermined by the experimenter (e.g., Pillutla & Murninghan, 1996; Sanfey et al., 2003). Alternatively, participants may interact over multiple iterations or rounds of an economic game with an ostensible partner, when in fact the partner's behavior is a computerized response determined by the decisions of the participant (e.g., a "tit-for-tat" strategy; van Lange, Ouwerkerk & Tazelaar, 2002). Using pseudo-dyadic tasks allows for non-interdependence in participant responses, enabling the testing of specific hypotheses with fewer participants. Additionally, conducting truly interactive studies over MTurk is quite complex and requires advanced programming and networking skills (Siddharth & Winter, 2012).

Given both the frequency of pseudo-interaction in the study of social behavior and the technological challenges involved in actually linking participants, our investigation focused on pseudo-interaction tasks in which there was never a real "partner" with whom a participant would interact. Instead, we used an online survey system (Qualtrics.com) that allows the presentation of screens to be timed, allowing us to lead participants to believe that they were waiting to be connected to and communicate with another person, when in reality the inputs from

the "partner" were pre-determined by the experimenters and the participants were in no way linked to another person. To ensure that participants would be debriefed about the true nature of the study (i.e., that there was no partner), we linked payment on the MTurk system to a "completion code" that participants received on the debriefing screen; so that participation was fully voluntary, participants were told that they could reach this screen without answering any questions.

To determine whether such pseudo-interactive behavior on MTurk is believable to participants and represents a reasonable approximation of that which researchers could expect in the lab, we conducted replications of well-established dyadic tasks. By examining tasks in which a clear and consistent effect has emerged in past research, we felt that null effects would be meaningful failures of MTurk as a venue for these tasks. Conversely, we believed that establishing reliable effects that descriptively parallel those in the lab would support the appropriateness of MTurk for these types of studies. We were therefore interested in establishing the qualitative assumption of interactivity by participants, rather than in conducting a full quantitative comparison to establish identical effect sizes and sampling distributions. In Study 1, we examined informational social influence by a supposed partner, which we compared to an anchoring effect from an estimate explicitly provided by the computer. In Studies 2-4, we raised the stakes for participants by examining behavior in economic tasks in which cooperation with a partner would lead the participant to forgo some amount of bonus payment. Finally, in Study 4, we also examined whether an established moderator of social behavior, ingroup versus outgroup status in a minimal groups paradigm, would likewise be able to influence social behavior on MTurk. Following the recommendations of Simmons, Nelson, & Simonsohn (2011) for minimizing the likelihood of false positive results, the studies below report all collected variables

and all experimental conditions; moreover, the reported studies represent all but one of the pseudo-dyadic studies we have attempted to date on MTurk. (We previously conducted a version of Study 2 with only one of the conditions, for which the effects were comparable.) Sample sizes were set *a priori* as noted in each study; the obtained *N* is slightly greater in Studies 1, 2, and 3 because some participants did not enter their payment code on MTurk, leading to additional "workers" being recruited by the automated system.

Study 1

Informational social influence, in which individuals alter their behavior because of the perceived value of the information provided by others, is one of the most classic social psychological phenomena (Deutsch & Gerard, 1955). Initially demonstrated in Sherif's (1935) study using the autokinetic effect, it has been shown to influence a wide range of decisions, from eyewitness recall in suspect lineups (Steblay, 1997) to product purchasing (Bearden & Etzel, 1982). Study 1 aimed to demonstrate this effect in participants on MTurk by comparing the effect of a low (versus high) anchor provided by a "partner" or by the computer on numeric estimates generated by participants. While even randomly generated anchors have an effect on numeric estimates (Tversky & Kahneman, 1974), the effect should be stronger if the anchor is believed to come from another person.

Method

102 participants (*a priori N* = 100) were recruited from Amazon's Mechanical Turk (MTurk) for a study with the description "Decision-making Study: This study examines how people make and react to decisions" for which they were paid 25 cents. Participation was limited to participants located in the United States with a past approval rating in the MTurk system above 95%.

Participants were told that they would be asked to complete an estimation task, and were randomly assigned to only the "computer anchor" or "partner estimate" condition. Participants in the "computer anchor" condition were informed that the computer would randomly generate a starting value for the estimation task. When participants advanced the screen, the message "Please wait while the computer generates a random number…" appeared on screen for 31 seconds.

Participants in the "partner estimate" condition were told that they would be paired with another participant, and that they would provide estimates sequentially, with the participant who went second being informed of the partner's estimate. When participants advanced the screen, the message "Please wait while the computer connects to another participant…" appeared for 8 seconds, followed by the message "Please wait while the computer assigns roles…" for 5 seconds. Participants were then informed they had been assigned to make the second estimate, and the message "Please wait while Player A completes the estimation task…" appeared onscreen for 31 seconds.

We used a slightly modified variant of a now-classic decision-making task introduced by Tversky and Kahneman (1974). All participants estimated the percent of UN member nations located in Africa[1] and additionally were told that because we were interested in the ways that people make numeric estimates, they should NOT look up the correct answer in another browser window. Participants were assigned to receive a low (10%) or high (65%) anchor. Participants first indicated whether the correct percentage was more or less than 10% (65%), and then provided an exact estimate.

Participants then completed a brief demographic questionnaire and a funnel debriefing asking what they thought the study was about and if anything had seemed odd or suspicious.

Participants were then provided with written debriefing information summarizing the nature and goals of the research and a code to enter on the MTurk website for payment. Participants in all of the following studies completed the demographic questionnaire and were debriefed and paid in a similar fashion. The results of the suspicion probe are discussed later.

Results and Discussion

Given the robustness of anchoring and adjustment in numeric estimation (Tversky & Kahneman, 1974), we expected that across both conditions, participants would make higher estimates in the high anchor condition than in the low anchor condition. However, to the extent that participants treated the partner as a real person, we further hypothesized that the anchor amount would interact with the source of the starting value (partner vs. computer) such that the anchoring effect would be larger in the partner estimate condition than in the computer anchor condition. A 2 (anchor value: high or low) X 2 (anchor source: partner or computer) ANOVA yielded support for both predictions. Our first hypothesis was supported by a significant main effect of anchor, $F(1,98) = 41.14$, $p < .001$. Higher estimates were made by participants who had responded to the higher anchor ($M = 32.13$, $SD = 16.02$) than the lower anchor ($M = 14.59$, $SD = 11.91$), see Figure 1. More importantly, the predicted interaction between anchor and source was significant, $F(1,98) = 6.22$, $p = .014$. Although significant in both cases, the effect size for the simple effect of anchor (high versus low) was nearly three times larger in the partner estimate condition ($d = 2.06$) than in the computer anchor condition ($d = .72$). Furthermore, the impact of a high or low anchor on participants in the "computer" conditions was extremely similar to that observed by Tversky and Kahneman (1974). In our study, estimates in the high anchor condition were 1.7 times larger than in the low anchor condition, whereas Tversky & Kahneman observed estimates 1.8 times larger. In contrast, participants in the "partner" conditions gave estimates

that were 3.18 times larger in the high anchor condition than the low anchor condition. Participants reacted to the starting values more strongly when they were ostensibly from another person as opposed to a computer program. These results provide initial evidence in support of the perceived plausibility of interacting with others in MTurk HITs.

Study 2

Study 1 demonstrated that participants were influenced by the belief that they were interacting with another person in the estimates that they provided following an anchor. However, there was no motivation for these participants *not* to behave as if the partner was real, so this provides a fairly weak test of whether participants on MTurk believe in such pseudo-interaction and therefore behave in similar ways as past participants interacting in lab studies. Study 2 therefore examined a context in which behaving as if the partner were real would come at a direct cost to the participant: the Dictator Game, an economic test of fairness (Forsythe, Horowitz, Savin, & Sefton, 1994). This game is used as a test of other-regarding decision-making, or altruism, by examining participants' willingness to share a sum of money with a partner. It is an important paradigm in that it avoids any strategic concerns about how a partner might respond to one's actions. Furthermore, the altruism displayed in dictator games cannot be explained by reputation or reciprocity concerns, as actors are typically informed that their partners will be unaware of their identities, and that they will not interact in the future; these instructions are analogous to the real constraints of MTurk, making the game additionally appropriate for the current research goals. Although identifying the dictator, the recipient, or both, increases allocations to one's partner, a substantial amount of altruism remains in the complete absence of identification (Frey & Bohnet, 1995, 1997). It is thus a critically important

paradigm in the behavioral economics and JDM literatures, and therefore important to validate for use in research on MTurk.

Study 2 compared participants who were told a partner would receive money based on their allocation decision (in a standard Dictator Game) to participants who believed that there would be no recipient of the portion of their allocation they chose to give away. Participants in the latter condition should be less likely to feel that fairness dictates they give some portion of the allocation away than participants who believe a real partner will receive money based on their allocation. However, if participants share the allocation due to a self-presentational or demand effect of not wishing to appear "greedy" to the experimenter, then these two groups should share identical amounts of the allocation.

Method

72 participants (*a priori N* = 70) were recruited from Amazon's Mechanical Turk as described in Study 1 and paid 30 cents (since this task took slightly longer than Study 1). Because we had previously run a version of Study 2 involving only the "partner" condition, we asked participants whether any part of the task seemed familiar. 23 participants reported having completed a task with some similarity to this research (interacting with a partner or dividing an allocation). We report results both for all participants and when these potentially non-naïve participants are excluded.

Participants were randomly assigned to one of only two conditions: the "partner" or "no recipient" condition. Participants were told that they would be connected to another participant, and that the computer would assign them to either allocate $3 in "bonus" funds or receive an allocation from another participant. When participants advanced the screen, the message "Please wait while the computer connects to another participant…" appeared for 8 seconds. The screen

then changed to the message "Please wait while the computer assigns roles…" for 5 seconds. Participants in the "partner" condition were then informed that they had been assigned the role of allocating the bonus funds and were asked to indicate how much to allocate to each player (the program would not allow them to advance unless the sum of the two values was exactly $3). They were informed that the funds would be distributed according to their allocations and that the other player would make no decision in the experiment. Participants in the "no recipient" condition were informed after the "please wait while the computer connects…" message that no other participant was available to be connected to, and they would therefore allocate money between themselves and "no recipient." They were informed that funds would be distributed according to their allocation even though there was no recipient for the portion of the allocation they gave away.

Results and Discussion

We were interested in whether participants who were told that they were interacting with another participant would share any of the allocation when doing so came at a direct cost to themselves. If participants in the "partner" condition actually believed that there was no other person and were simply complying with an implied demand characteristic, then they should give the same amount of money away as participants in the "no recipient" control condition. However, consistent with a belief that their partner was real, a substantial majority of participants in the "partner" condition shared some amount of the allocation (75%). In contrast, only 30.5% of "no recipient" participants did so, $\chi^2(1) = 14.27$, $p < .001$. This suggests that a demand effect alone was not responsible for sharing. (For naïve participants only, 77% of "partner" and 30% of "no recipient" participants shared the allocation, $\chi^2(1) = 10.66$, $p < .001$).

We also examined how much participants would allocate to the recipient. As expected

participants in the "partner" condition gave significantly more of the $3 allocation ($M$ = $0.97, SD = 0.62) to the recipient than participants in the "no recipient" condition (M = $0.36, SD = 0.56), t (69) = 4.29, p < .001. (For naïve participants only, M = $0.95, SD = .61 & M = $0.37, SD = .59, respectively, t (47) = 3.40, p < .001). Engel (2011) conducted a meta-analysis of dictator games conducted in laboratory settings and found that the average rate of giving in those studies was 28%. The average percentage of the allocation given away by participants in the "partner" condition (32%) is comparable to this meta-analytic benchmark value of 28% (i.e., $0.84), one-sample t (35) = 1.19, p = .24, (for naïve participants only, t (25) = 0.93, p = .36). In contrast, participants in the no recipient condition shared far less of the allocation (12%) than is typical in lab studies, one-sample t (34) = 5.08, p < .001, (for naïve participants only, t (21) = 3.57, p = .002).

Participants in the "partner" condition thus appeared to behave as if interacting with a real partner even when doing so came at a direct financial cost, and did so in a manner descriptively comparable to participants in laboratory dictator games (Forsythe, 1994; Engel, 2011). In contrast, participants in a control condition were less likely to share the allocation than those in the "partner" condition, and shared far less than is typical of lab participants, suggesting that the effects in the "partner" condition are not simply due to a demand effect.

Study 3

Study 2 established that participants would behave as if a partner were real by sharing an allocation of bonus funds even though doing so came at a cost to themselves. Study 3 aimed to replicate this effect in a different economic game, the Ultimatum Bargaining Game (UBG), in which participants chose whether to reject an unequal allocation that nevertheless offered a small reward for accepting (Guth, Schmittberger & Schwarze, 1982). This paradigm is commonly

used in JDM studies of interdependent decision-making; for instance, a meta-analysis by Oosterbeek, Sloof, van de Kuilen (2004) included 37 published papers using this paradigm. It has been the recipient of this considerable attention because it is a strong demonstration of behavioral data diverging from game theoretic, or strictly "rational" decision making. The fact that respondents are willing to reject small offers from proposers suggests that much more than strictly economic concerns are at play, such as an aversion to unfairness (Thaler, 1988).

Specifically, Study 3 compared participants who were told a partner made this unequal allocation to participants who believed that a computer had randomly determined the allocation. Participants in the latter condition should be less likely to feel that fairness has been violated than participants who believe a real partner made an unfair allocation, and thus more likely to accept the offer.

Method

61 participants (*a priori* $N = 60$) were recruited and paid as described in Study 2. Participants were randomly assigned to one of only two conditions: the "partner" or "computer" condition.

Participants in the "partner" condition received the same instructions as the "partner" condition in Study 2, except they were told that there would be $5 of bonus funds allocated and that the "recipient" would be able to reject the offer, in which case neither participant would receive anything. After seeing the "connecting" and "assigning" screens as in the "partner" condition in Study 2, participants were told they had been assigned to be the recipient, and told to wait while the other participant made an offer. This message appeared onscreen for 31 seconds, after which the computer automatically advanced. Participants then saw a message that their partner had allocated $4.75 to him- or herself and $0.25 to the participant. They were asked

to accept the offer or reject it, and reminded that if they rejected it, neither participant would receive a bonus.

Participants in the "computer" condition were given the same instructions except they were told the computer, rather than another participant, would be their partner. They did not see the "connecting" message, but the procedure was otherwise identical to the "partner" condition.

Results and Discussion

As predicted, participants who were informed that a partner made the allocation were less likely to accept the offer than those who were told the distribution was generated by the computer. 41.9% of participants in the "partner" condition accepted the unfair offer. In contrast, 76.7% of participants in the "computer" condition accepted the unfair offer. These two frequencies differed significantly, $\chi^2(1) = 7.60$, $p = .006$, indicating that participants were substantially less likely to accept an unfair offer ostensibly from another participant than from a computer.

These results descriptively parallel a similar study conducted in the lab (van 't Wout, Kahn, Sanfey, & Aleman, 2006). Although overall acceptance rates of very unfair offers were lower in their study, at 20% for human offers and 38.3% for computer offers in their study versus 41.9% and 76.7% in the current study, the effect of the source of the offer was quite similar to the present study. Both the present study and van't Wout et al. (2006) found that unfair offers from human partners are roughly half as likely to be accepted as identical offers from computer partners: in their study, unfair offers were 55% as likely to be accepted when coming from a person than a computer, compared to 52% in the current study.

Study 3 thus offers further support for the use of MTurk for pseudo-interactive tasks. Participants in the "partner" condition behaved as if they were interacting with another person,

refusing to accept an unfair offer even when doing so caused them to forego a monetary reward. Furthermore, MTurk participants behaved similarly to a lab sample in the relative likelihood of rejecting an unfair offer from a partner versus a computer.

Study 4

Study 2 demonstrated that MTurk participants shared an allocation with a partner at a much greater rate than in a control condition, indicating that they behaved as if the partner were real even when this imposed a cost. Study 3 demonstrated that participants responded differently to a computer than a partner in an ultimatum game, although this again came with a cost. These studies thus provide substantial support for the feasibility of doing pseudo-interactive tasks on MTurk. However, it is also important to determine whether factors that influence social behavior in the lab also influence behavior on MTurk.

Study 4 thus examined whether an important social moderator, ingroup versus outgroup membership, would likewise moderate the amount shared in a dictator game. Individuals generally show a preference for and favorable treatment of those who are members of their own groups relative to those who do not share these memberships, known as ingroup favoritism. This important phenomenon is well-documented within social psychology (for instance, a 1992 meta-analysis on the effect by Mullen, Brown, and Smith included 137 studies). This phenomenon addresses issues of major social significance and is at the center of a rich theoretical literature. In particular, the minimal groups paradigm developed by Tajfel and colleagues (1971) likewise has a long history within the field. In the minimal groups paradigm, individuals are categorized into one of two meaningless groups. Participants generally show favoritism for others sharing the same designation, an important demonstration of ingroup favoritism given the subtlety of the manipulation and lack of meaningful reasons for such favoritism. Finding evidence of ingroup

favoritism using the minimal groups manipulation in online pseudo-interaction would offer important support for the use of MTurk for such research, as it would demonstrate that an important but subtle moderator of social interaction in the lab also affects behavior on MTurk.

Method

75 participants (*a priori N* = 75) were recruited as described in Study 2 and were paid 25 cents for their participation. Participants first were asked to perform a dot estimation task in which an array of dots was presented on-screen for 10 seconds, after which participants were asked to estimate the number of dots in the array. After providing their estimate, participants were informed that people have general tendencies to over- or under-estimate numeric quantities in various contexts, and were randomly assigned to receive feedback that their response indicated that they were an Overestimator (Underestimator).

25 participants were assigned to each one of only three conditions: computer partner, ingroup partner (sharing the same dot estimator classification), or outgroup partner (having an opposite classification). All were given the same Dictator Game instructions and assignment ("Player A") as in Study 2. Specifically, all participants were told that they would either allocate $5 in "bonus" funds or receive an allocation from their partner. Participants in the computer partner condition were additionally informed that their "partner" would be the computer.

When participants advanced the screen, the message "Please wait while the computer connects to another participant…" appeared for 8 seconds for those in the ingroup and outgroup partner conditions. All participants then saw the message "Please wait while the computer assigns roles…" for 5 seconds. All participants were informed that they had been assigned to allocate the bonus funds.

Participants in the ingroup and outgroup conditions were then told "To protect the

anonymity of all participants we cannot give you any identifying information about your partner. However, since most participants prefer to have some information about their partner, we can inform you of your partner's categorization based on the dot estimation task. Your partner is a dot (overestimator/underestimator)."

All participants were then asked to indicate how much to allocate to each player (as in Study 2, the program would not allow them to advance unless the sum of the two values was exactly $5). They were informed that the funds would be distributed according to their allocations and that the other player would make no decision in the experiment.

Results and Discussion

As in Study 2, the key dependent measure was the amount participants would allocate to the other player. A one-way ANOVA revealed an effect of condition on allocations, $F(2,75) = 3.90$, $p < .05$. Participants in the ingroup condition gave significantly more of the $5 allocation to their partner (M = \$2.23, SD = 0.83), than those in the outgroup condition (M = \$1.64, SD = 1.14), $t(48) = 2.14$, $p = .037$ and those in the computer partner condition (M = \$1.52, SD = 1.04), $t(51) = 2.77$, $p = .008$, which did not differ, $t(45) = .37$, $p = .72$. As would be expected based on past lab research, minimal group status of an online partner affected the amount that participants shared with these partners, with participants sharing the most with an ingroup partner. Furthermore, those in the ingroup condition gave a greater percentage of the allocation than the meta-analytic baseline rate of giving in laboratory Dictator Games (28%, i.e. \$1.42; Engel, 2011), one-sample t (27) = 5.20, $p < .001$, whereas the giving of those in the outgroup (one-sample t (21) = 0.90, $p = .38$) and computer (one-sample t (24) = 0.48, $p = .63$) conditions did not differ from the typical rate in previous lab tasks (for which no group membership was specified).

It may seem surprising that the computer and outgroup conditions did not differ. However, we believe at least two explanations could account for this effect. First, one of the major findings in the literature on intergroup relations is outgroup dehumanization, in which members of outgroups are considered less than fully human. Along with *animalistic dehumanization*, in which outgroup members are conceptualized as non-human animals, dehumanization may take the form of *mechanistic dehumanization*, in which outgroup members are considered akin to robots or computers (Haslam, 2006; Haslam, Kashima, Loughnan, Shi, & Suitner, 2008). Thus, the equivalence of outgroups and the computer may be evidence of mechanistic dehumanization of the outgroup partner. Alternatively, it may be the case that the "computer" was imbued with some human or social qualities. Given that MTurk is presented as a "marketplace," the "computer" may have been understood to be the researcher ("requester" in MTurk's terminology). Participants ("workers") could thus have felt social pressure to avoid being seen as greedy in the amount they kept, as with the small number of participants in Study 2 who gave money to "no recipient." In the MTurk system, requesters can give "qualifications" to workers to make them eligible for further work. Likewise, the rate of approval by past requesters is a prerequisite to be eligible for certain tasks (e.g., in the current research, we only allowed workers with a past approval of 95% or more to participate). There are thus real impression management concerns for workers in their interaction with the "requester" (which, we would note, raises ethical concerns about the potential for coercive behavior by experimenters). In short, we believe that the equivalence of these two conditions is not wholly unsurprising, and more importantly, we believe that it does not weaken our core assertion that participants in the ingroup and outgroup conditions generally behaved as if their partners were real.

Meta-Analysis: Rates and effects of suspicion

In all four studies, participants completed a funnel debriefing. Participants were first asked what they thought the study was about. Across all studies, only 1 participant (of 310 total participants) expressed suspicion that the partner was real at this stage. Participants were then asked if anything about the study had seemed "odd or suspicious." Table 1 shows the percent of participants in conditions with a "partner" who mentioned that they felt there was something odd about having a partner (which could include, for instance, the normative observation that this is an unusual feature of MTurk tasks or a comment about the partner's selfish behavior in Study 3) or explicitly voiced suspicion that their partner was not a real person. Overall, 21.3% of participants expressed specific suspicion in the funnel debriefing. However, this belief had a fairly limited influence on their behavior. As shown in Table 1, the effect sizes among those voicing suspicion and those not voicing suspicion were generally comparable. In Study 1, the effect was weaker for the suspicious than the unsuspicious participants (though still categorized as a "large" effect under Cohen's guidelines for the interpretation of *d*). In Studies 2 and 3, the effect was stronger for suspicious participants, which meant they were actually more likely to incur a cost despite their disbelief. In Study 4, the effect size was weaker for suspicious participants. Given that there was no consistent effect of suspicion and only 1 participant reported suspicion about their partner without direct prompting, we believe that suspicion poses an acceptable risk given the benefit of being able to conduct pseudo-dyadic tasks on MTurk.

General Discussion

Across four studies, we demonstrated that participants on MTurk responded to a "partner" as if they were real and in a qualitatively similar way as laboratory participants in previous research. Participants used information provided by the partner more heavily than a computer-generated estimate in providing a numeric estimate, shared a real allocation of money,

reacted to unfairness in an offer made by a partner by forgoing a small reward, and showed sensitivity to the minimal group membership of a partner. These data provide support for the potential to use MTurk for pseudo-interactive research, increasing the value of this tool to psychologists interested in social and interdependent phenomena, such as social psychologists and JDM researchers.

Of course, there are several important caveats regarding this research. One is that participants did occasionally voice suspicion about their partner during a funnel debriefing. However, as noted in the meta-analysis, this suspicion had little consistent effect on behavior. It was also least common when the presence of the partner was well justified but not obviously central to the research question, as in the anchoring study. In tasks overtly concerned with cooperation, participants were clearly aware of this purpose of the research and seemed increasingly suspicious as a result. The cover story therefore seems to be particularly important to the success of pseudo-dyadic research. Additionally, our comparison to past research was primarily qualitative in nature, and we can offer no conclusions about whether the sampling distributions obtained on MTurk are different in any important ways from those that we might have obtained in the lab. Furthermore, the current studies do not capture the full range of social tasks that researchers may be interested in. In particular, participants received no direct communication from their partners (other than the numeric estimate in Study 1). Although many studies of social and interdependent behavior do not involve such communication, the present research cannot speak to the feasibility of research that does involve richer "interactions." Likewise, we only studied ostensible dyads. It is possible that larger groups would not produce effects, particularly if the size begins to strain credibility – for instance, a participant may be unlikely to believe that she has been connected to 10 other participants simultaneously within a

matter of seconds. Finally, we wish to emphasize the ethical concerns posed by internet research. Given that participants are "workers" dependent on the approval of the experimenter for both payment in the current study and eligibility for future tasks, there is a high risk for coercion that researchers need to mitigate. In the current research, we did so by informing participants that they could skip any question they did not wish to answer and still obtain the completion code, and by avoiding, as much as possible, the use of settings in the web-based system that would require participants to provide a response to continue. Furthermore, unlike in the lab, there is no guaranteed way to debrief participants who choose to cease participation before the end of the study, as participants can simply close their web browser to cease participation. In the current research, our payment system was a relatively successful attempt to reduce this possibility. Across all four studies, 93% of participants who began the study completed it through the debriefing. However, studies in which a lack of complete debriefing could pose a substantial risk of harm (e.g., studies in which the partner rejects, criticizes, or insults the participant) may never be appropriate for MTurk. Although these risks are not unique to pseudo-interactive tasks (e.g., there could also be potential harm from false feedback on an online personality measure) they may be heightened in social interactions.

Overall, these studies provide support for the appropriateness and plausibility of conducting pseudo-interactive tasks on MTurk. We feel that this research offers an important extension of methodology using MTurk, and hope that others find it useful in expanding the range of online research.

Note:

[1] At the time this study was conducted, there were 54 nations in the African Regional Group of the United Nations, 28% of the 193 member nations (United Nations, 2012).

References

Bearden, W.O., & Etzel, M.J. (1982). Reference group influence on product and brand purchase decisions. *Journal of Consumer Research*, *9*. 183-194.

Blackhart, G. C., Nelson, B. C., Knowles, M. L., & Baumeister, R. F. (2009). Rejection elicits emotional reactions but neither causes immediate distress nor lowers self-esteem: A meta-analytic review of 192 studies on social exclusion. *Personality and Social Psychology Review, 13,* 269-309.

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.

Deutsch, M., & Gerard, H. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*, 629-636.

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

Forsythe, R., Horowitz, J.L., Savin, N.E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior, 6*, 347-369.

Frey, B., & Bohnet, I. (1995). Institutions affect fairness. *Journal of Institutional and Theoretical Economics*, *151*, 286-303.

Frey, B., & Bohnet, I. (1997). Identification in democratic society. *Journal of Socio-Economic,s 26*, 25-38.

Guth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental-analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367-388.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10,* 252-264.

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, inhuman, and superhuman: contrasting humans with nonhumans in three cultures. *Social Cognition*, *26*, 248-258.

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

Mullen, B., Brown R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*, 103-122.

Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, *7*, 171-188.

Pillutla, M.M., & Murnighan, J.K. (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes*, *68*, 208-224.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments using Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411-419.

Rand, D. G. (2011) The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology,* DOI: 10.1016/j.jtbi.2011.03.004.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., & Cohen, J.D., (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, *300*, 1755-1758.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *27,* No. 187.

Siddharth, S., & Winter, M. (January, 2012). *Conducting synchronous experiments on mechanical turk.* Paper presented at Society for Personality and Social Psychology, San Diego, CA.

Simmons, J.P., Nelson L.D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359-1366.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *25*, 459-473.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology, 1,* 149-178.

Thaler, R.H. (1988). Anomalies: The ultimatum game. *Journal of Economic Perspectives*. 2, 195–206.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124-1131.

United Nations (2012, March 6). Department for General Assembly and Conference Management: United Nations Regional Groups of Member States. Retrieved from http://www.un.org/depts/DGACM/RegionalGroups.shtml.

Van Lange, P.A.M., Ouwerkerk, J.W., & Tazelaar, M.J.A. (2002). How to overcome the detrimental effects of noise in social interaction: The benefits of generosity. *Journal of Personality and Social Psychology*, *82*, 768-780.

van 't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research*, 169, 564-568.

Williams, K.D., Cheung, C.K.T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology*, *79*, 748-762.

Table 1

Study	% initially suspicious	% mentioning partner in probe in any way	% suspicious in probe	Effect for unsuspicious	Effect for suspicious
1	0%	23%	15%	**High vs. Low *d* = 2.13**	High vs. Low *d* = 1.39
2	0%	22%	14%	Mean value *d* = 1.49	**Mean value *d* = 1.79**
3	3%	35%	25%	Percent rejecting = 52%	**Percent rejecting = 75%**
4	0%	26%	24%	**Ingroup vs. outgroup *d* = 0.68**	Ingroup vs. outgroup *d* = 0.17

Note: Data for Study 2 includes all participants in the "partner" condition. "Initially suspicious" refers to participants who voiced suspicion about the partner when asked what they thought the study was about. "Mentioning partner" refers to participants who made some reference to the partner when asked if anything about the study seemed odd; "suspicious in probe" are participants who explicitly voiced suspicion that the partner did not exist in response to this question. The "effect for suspicious" effect sizes are based on this set of participants. The larger effect size (unsuspicious vs. suspicious participants) is shown in **bold.**

Figure 1

Mean estimates provided in Study 1 as a function of whether the anchor was high (65%) or low (10%) and purportedly generated by another person or the computer.