Does being observed enhance self-control?

Florian Diekert and Kjell Arne Brekke: December 2, 2019

This study plan is an extension of the study plan with the same title from May 13th 2019. We planned and conducted a study on MTurk. The results did not confirm our hypotheses, but we also find reason to believe that these results were contaminated by a message on the MTurk message board posted on the first day of our main data-collection. The comment read "fun game. Bonus includes your points and your partners, my bonus for the HIT only includes mine as the rest is calculated." As the main treatment difference was whether the partners score was real or calculated, we fear that this contaminates the study.

A further concern came up as we looked deeper into the data. We collected some observation from one treatment prior to the main data-collection, to test some server issues with the program. Looking at the main outcome variable for these two days they are significantly different from each other and the difference is much larger than any treatment effect from the later data collection. This indicate that there is too much noise with this kind of experiment on MTurk, to really trust the data.

The new study will redo the experiment with a different sample, but with the same main hypotheses and with the same basic design. There will be only a few changes. One difference to the previous study is that we remove the "AO" treatment. In this treatment players are matched with a computer – the partner score is calculated. But they are still observed by someone. As there is no partner to observe them, they will be observed by a random player with whom they have no joint interest. We had a concern with the previous study that this would appear odd, and the data indicate that the treatment stands out in the amount they borrow from their partner, but in a way that does not fit in with previous literature. We do not see a plausible mechanism behind this, except that the instructions have appeared odd, we thus decide not to pursue this further and drop the treatment, and separate the difference between observability and joint interests as explained below.

Design

We follow the main features of the design in our previous study (Diekert and Brekke, 2019) which again is based on Shah (2012).[1] Participants are recruited through Amazon Mechanical Turk to play a guessing game known from the TV-show Family Feud. The participants will guess the five most popular answers to questions such as "Name five things you take on a picnic". Participants are matched in pairs and paid for each correct answer they, or their partner, give.

[1] Details on experimental design of our previous study can be found in the corresponding study plan: https://www.socialscienceregistry.org/trials/2582/history/23269

Initially, participants have 15 seconds per questions, but they can borrow time from future questions. This has a 100% interest rate, thus each second used takes two seconds from the time budget for future questions. The first player in the pair can further, when their time is exhausted, take time from their partner. The same 100% interest rate applies; each second they spend after their budget is exhausted will reduce the partner's budget with two seconds.

The treatments differ along the dimensions: **observability** and **group composition**. In all treatments, the first player can take time from the second player with a 100% interest rate. For **group composition,** the two alternatives are either *Pair (P)*, where there are two players in a sequence, or *Alone (A),* where the second player is a computer that will score like a real player given the same amount of remaining time. For **observability**, the two alternatives are either *Observable (O)* or *Non-observable (NO).* In the observable treatments, the first player will be told that the second player will be informed about the first players borrowing behavior. The second player will be told about total borrowing in round 1-4 as well as total borrowing in all rounds, and how much time the player has taken from the second player. As borrowing in rounds 1-4 is a key outcome variable we will not highlight this unequally across treatment, thus instruction to first player will only state second players will learn about their borrowing behavior. The previous study used a 2 x 2 design, but as pointed out above the *Alone* x *Observable (AO)* treatment had very strange results. In that treatment the second player is not in the same pair, but a player who plays the same game after the first player has completed. We fear that it seems odd to be observed by some random player, so we have decided against including this treatment in the current study.

The design is thus as follows:

Treatment	**Description**
PO: Pair & observable	A real partner will be informed about first players borrowing behavior
PNO: Pair & not observable	A real partner that will be informed only about own time budget
A: Alone (& not observable)	No real partner, second player score is calculated

Incentives

We originally pay 0.5$ for participation and 0.10$ for each correct answer. We calculated that this would give at least the minimum wage of 12$ per hour, while most participant would earn more. For Norwegian participants we must pay more, at least 250 Kr per hour.

We expect the survey to take 10-15 minutes, and hence we should pay at least 65 kroner per 15 minutes. The average score for MTurkers is 20-26, but Norwegian participants would know less about what an American sample answered, so score will be lower. Moreover, few participants will know the game family feud, which also lower expected score. With 20 correct answers and 20 kroner for participation, we should pay 45/20=2.25 kroner per correct answer. To account for lower score we could set it to 3 kroner per correct answer. To enhance participation, and as it seems reasonable with a

higher hourly wage for a 10-15 minutes job, we increase show up fees to 40 kroner and keep payment at 3 kroner per correct guess.

Definition – key variables

This section is identical to the previous study plan.

The main hypothesis in this project is about the amount of self-control for the first player. Consider first the hypothetical case where the player cannot take time from the second player. The results both the results from Shah et al (2012) as well as our own initial results show that the optimal time use is to use 15 second on each question. Any borrowing from the players own time budget is inefficient. The temptation to borrow is largest at the beginning, then there is much time left. Thus, a possible measure of self-control is the borrowing in the first rounds. In the instructions, we focus in particular on the first four rounds, so total borrowing in the first four rounds would be a measure. The next question is how to deal with the possibility to take from the second player.

With the possibility to take from a second player, there are two possible motives for borrowing in the first rounds. (1) The player borrows with the intention to take time from the partner. (2) The player borrows from a lack of self-control. For case (1), consider a player who decides to borrow half the time budget of the partner, but otherwise have perfect self-control and is able to spend the exact same time on all rounds. This would imply spending about 16 seconds per round, rather than 15 seconds. This amount of borrowing in the first four rounds is very small relative to the amount of borrowing observed in the corresponding treatment of our previous study (there, participants borrowed 20.6 seconds, difference is strongly significant, $p<0.001$). Note also that the first player will first borrow from his own future time budget. The message that he is borrowing from his partner will only appear when his own time is entirely exhausted. We thus think that the total amount of borrowing in the first four rounds is a good measure of self-control, and that variations in the intention to take from the partner will represent negligible noise, even if it should correlate with the treatment.

Sample Restriction:

In the previous m-turk sessions we observed that some first players let the clock run and exhausted the entire time budget in the first round. We exclude all observations from participants that in the first round exhaust their entire time budget, including what they can borrow from their partner. In addition, we exclude observations from players where not all rounds results are recorded.

Main outcome variable: $B_4\sim$ The remaining time budget after 4 rounds (=initial time budget minus what has been used in the first four rounds (at the task or for paying interest), the lower the more time a participant borrows). Measure of self-control; a higher value of B_4 is better self-control.)

The following variables are included as control-variables in potential regressions on the main effect.

Total time budgeting: $B_{12}\sim$ The remaining time budget after 12 rounds (=initial time budget minus what has been used in the twelve rounds of the game (at the task or for paying interest), the lower the more time a participant borrows, a negative value for the first player shows how much he/she borrowed from the second player, B_{12} cannot be negative for second players).

Score; individual and pair: $S_i\sim$ The score of player *i=1,2;* number between 0 (when no answer to no question has been found) and 60 (when all answers to all questions have been found). $S\sim$ Score for the pair. $\hat{S}\sim$ Score for the pair when we use calculated score for the second player.

Gender: G

Age: A

Know Family Feud: K

Pair: $P\sim$ Dummy for solitude treatments ($P = 0$ for alone, $P = 1$ for pairs.)

Observed: $O\sim$ Dummy for treatments where later players observe borrowing. ($O = 1$ when observed, $O = 0$ otherwise.)

Hypotheses

The main hypothesis is that the combination of joint interest and observability should cause less borrowing in the first four rounds. We further believe that both joint interest with a real person and being observed contribute to the difference, but we do not know what contribute most. Thus, we expect the treatment PNO to be in-between the two other treatments.

Hypothesis A. $E(B_4|PO) < E(B_4|A)$

Hypothesis B. $E(B_4|PO) \leq E(B_4|PNO) \leq E(B_4|A)$

Reconsidering previous results.

In our first study we could not reject a claim pairs had the same score irrespective of whether they had joint time budget. This was surprising first players did borrow from their partner when they could, and borrowing was obviously harmful. However, with no real partner they performed much worse. This difference in performance was explained by the fact that first players with a joint budget but with a real partner borrowed much less in the beginning of the game, than those with no real partner. A natural hypothesis is that this will still be the case.

Hypothesis C. $E(\hat{S}|PO) > E(\hat{S}|A)$

Tests and regressions

We will test all hypotheses in pairwise t-tests. Note that some hypothesis state weak inequalities, in those cases we should not be able to reject a hypothesis of equality in a one-sided test.

Power calculations

In the current study we don't expect to be able to recruit more than 100 per group at most, that will require 500 participants. The invitation will be sent out to a list of 1300 E-mails. As the experiment can be done online at a suitable time and with a good payment, a 40% sign up rate is not unreasonable. To ensure enough power on the first hypothesis, hypothesis A, we will send out the invitation to 300 E-mail first, where we include all 3 treatments. If the response rate is below 30%, we will include only treatment A and PO in the invitation to the remaining 1000 participants.

The previous plan indicated that we should have more than 85% power with 200 observations in each group. On the other hand, the experiences from the failed study also indicated that data from MTurk is very noisy for this kind of experiment, and thus we expect a considerable reduction in standard-deviations. If standard-deviations are reduced by a factor of 2, a sample size of 50 in each group should be enough to get 85% power. While we cannot do a more exact power analysis, it seems reasonable to expect that the reduction in SD and reduction in sample size about cancel out, and that we maintain an acceptable power.

References

Charness, G., & Sutter, M. (2012). Groups make better self-interested decisions. Journal of Economic Perspectives, 26(3), 157-76.

Diekert and Brekke (2019): Groups discipline resource use under scarcity. Unpublished

Falk, A., & Ichino, A. (2006). Clean evidence on peer effects. Journal of labor economics, 24(1), 39-57.

Herbst, D., & Mas, A. (2015). Peer effects on worker output in the laboratory generalize to the field. Science, 350(6260), 545-549.

Shah, A. K., Mullainathan, S., and Shafir, E. (2012). Some consequences of having too little. *Science*, 338(6107):682-685.

Zajonc, R. B. (1965). Social facilitation. Science, 149(3681), 269-274.