**Intuitive Prosociality:
Heterogeneous Treatment Effects or False Positive?***

Eirik Strømmland[†]

Department of Economics, University of Bergen
Hermann Fossgate 6, 5007 Bergen, Norway

Gaute Torsvik

Department of Economics, University of Oslo
Moltke Moes vei 31, Eilert Sundts Hus, 0317 Oslo, Norway

Abstract

Heterogeneous treatment effects make it difficult to extrapolate from one research setting to another. However, what appears to be differences in effects across subpopulations may simply be false positives. This paper uses a representative sample of the Norwegian population (N = 1390) to systematically test for several proposed sources of heterogeneity in the literature on intuitive prosociality – a literature with large variation in results, which some researchers claim results from heterogeneity in the underlying effect. We use time pressure to induce intuitive decision making, and exogenously vary participants' experience with the game. We find no overall effect of time constraints on dictator game for inexperienced subjects, and there is no evidence for an interaction effect between subject experience and the effect of time pressure. As a more general test of treatment effect heterogeneity, we consider the full distribution of treatment effects conditional on various proposed moderators in the literature. The distribution of conditional effects is consistent with no causal effect of time pressure on giving and no systematic heterogeneity in the underlying effect across subpopulations.

* We thank Magnus Johannesson and Matthias Sutter for comments.

[†] Corresponding author. E-mail: Eirik.Stromland@uib.no

1. Introduction

A fundamental question in empirical economics is to what extent treatment effects generalize from one setting to another. If treatment effects vary a lot across populations, it is difficult to draw general conclusions about behavior from one study alone. Systematic variation in treatment effects will also make it more difficult to interpret replication failures, as a replication failure could simply result from the data being collected in a different population or context.

Researchers often investigate heterogeneity by splitting their sample by subgroups and test for differences in treatment effects across subgroups. An issue with this procedure is that if statistically significant effects are more likely to be reported and published, a research literature may consist of a pattern of mixed findings even though there is no heterogeneity in the effect of interest. Thus, researchers may mistake false positive findings for heterogeneous effects.

This paper tests for the presence of heterogeneous treatment effects in the literature on intuitive prosociality. Much of this literature is motivated by the “Social Heuristics Hypothesis” (SHH) which posits that intuitive decisions draw upon heuristics that have evolved through everyday experience where prosocial behavior is typically advantageous (Bear & Rand, 2016; Rand et al., 2014). Whereas some studies suggest a link between intuition and cooperation (Døssing, Piovesan & Engström, 2017; Everett et al., 2017; Rand, Greene & Nowak, 2012; Rand, 2016), several studies have failed to find a positive effect of time pressure or other manipulations of the level of deliberation that lies behind decisions (Tinghög et al., 2013; Verkoeijen & Bouwmeester, 2014; Bouwmeester et al., 2017; Camerer et al., 2018). Similar treatments in the Dictator game also contain mixed evidence; some find no effect (Hauge et al., 2016; Merkel & Lohse, 2018; Tinghög et al., 2016) and others report an effect (Schulz et al., 2014), while a meta-analysis suggests that there is an effect of intuitive processing on altruism only for women (Rand et al., 2016). A large string of papers has suggested that the effect of intuitive decisions on prosocial behavior is heterogeneous (Alós-Ferrer & Garagnani, 2018;

Chen & Krajbich, 2018; Rand et al., 2014; Rand, 2016; Rand et al., 2016; Rand, 2018; McAuliffe et al., 2018; Strømmland, Tjøtta & Torsvik, 2016; Goeschl & Lohse, 2018), which could account for the mixed findings in this literature.

We test for an effect of intuitive decision processes on altruistic behavior in a Dictator game in a large representative sample of the Norwegian population ($N = 1390$). We conduct two tests for heterogeneity in the underlying effect of time constraints. First, we causally manipulate subjects' experience by randomly assigning a random part of our sample to participate in a standard Dictator game without time constraints prior to participating in the Dictator game with a time constraint treatment. Importantly, no prior experimental games had previously been conducted on this recruitment platform, indicating that our sample should be highly inexperienced with experimental games. Previous studies have suggested based on correlational evidence that experience undermines the effect of intuition manipulations and that the SHH predicts this pattern as experienced subjects may have already adapted their heuristics to the situation (Rand et al., 2014). This hypothesis was also recently used by Rand (2018) to explain a replication failure by Camerer et al. (2018). Another reason to study the causal effect of experience on the time pressure effect is that experience seems to be negatively correlated with treatment effects in general (Chandler et al., 2015).

In a more general test for heterogeneous treatment effects, we consider the distribution of treatment effects conditional on a large set of variables the literature suggest may moderate the effect of time pressure on prosocial behavior. This analysis approach is related to statistical techniques that use the distribution of effects rather than single results to draw conclusions (Leamer, 1983; Sala-I-Martin, 1997; Simonsohn, Nelson & Simmons, 2014; Steegen et al., 2016). If there is no underlying effect and there is no genuine heterogeneity in the effect, then we should expect a small fraction of statistically significant findings in our conditional distribution of treatment effects. We find no evidence for heterogeneity when examining this

full distribution of conditional treatment effects, in contrast to the hypothesis that the inconclusive evidence of intuition effects may be explained by underlying heterogeneity in how intuition affects prosociality. We find no evidence that the time constraint effect depends on subjects' predisposition towards prosocial behavior (Alós-Ferrer & Garagnani, 2018; Chen & Krajbich, 2018). Neither do we find support for the hypothesis that intuitive decisions are prosocial only for women (Rand et al., 2016), or that subjects' level of trust in everyday institutions plays a role (Rand & Kraft-Todd, 2014).

Our findings are consistent with a theoretical model by Maniadis, Tufano & List (2014; 2017) building on Ioannidis (2005), where an increase in researcher degrees of freedom – the number of possible statistical tests that may be used to support a given alternative hypothesis - lowers the post-study probability that a randomly chosen statistically significant test is true. That model suggests that literatures where the research hypothesis is frequently supported by statistically significant subgroup tests are more likely to contain false positive findings. The statistical approach we adopt in this paper may be a valuable tool to avoid cumulation of errors in a research literature that leads to the impression of heterogeneous treatment effects.

This paper proceeds as follows. First, we present the details of the experimental design and the recruitment platform used to sample the study participants and outline our procedure for testing for treatment effect heterogeneity. Second, we analyze and discuss the experimental data. Finally, we conclude the paper.

2. Design and Methods

2.1. Experimental Design

For several reasons, we choose to study the effect of time constraints in a Dictator game rather than the conventional Public goods game (Rand et al., 2012). First, the Dictator game abstracts from strategic interaction, which enables us to isolate the effect of a given cognitive processing

manipulation on social preferences. Second, due to the element of strategic interaction the Public goods game is more difficult to understand, and confusion has been suggested to confound or moderate the effect of time constraints on behavior (Everett et al., 2017; Goeschl & Lohse, 2018; Strømmland et al., 2016). By studying a Dictator game, confusion is unlikely to be a possible source of variation in the treatment effect. Finally, even though some suggest that the SHH predicts an effect of intuition on altruism only for women (Rand et al., 2016), our interpretation of the SHH is that subjects acquire prosocial heuristics from experience with social dilemma games experienced in everyday life and that these heuristics are used in unfamiliar situations if the decision must be made without much time to reflect. It seems that an intuitive mindset will not be able to distinguish between different forms of sharing decisions (e.g. dictator game or prisoner's dilemma), which means that the SHH predicts that intuitive decision making will increase sharing both in the dictator game and intuitive cooperation in prisoner's dilemmas. In fact, empirical evidence seems to support this idea as it has been shown that experience with cooperative behavior leads to a subsequent increase in both dictator game giving and cooperative behavior in one-shot games (Peysakhovich & Rand, 2016).

A valid test of the Social Heuristics Hypothesis requires subjects who are inexperienced with the type of “laboratory games” used to test for cooperation and prosociality. The reason is that subjects unfamiliar with such games will draw upon heuristics derived from non-anonymous repeated games played in their everyday lives, whereas experienced subjects already have adapted to the atypical one-shot anonymous setting (Bear & Rand, 2016; Rand et al., 2014). To assure prior inexperience to lab-games, we decided to use a novel recruitment platform, the Norwegian Citizen Panel (NCP).

The Norwegian Citizen Panel is designed to obtain a representative sample of the Norwegian population. First, a sample of 25000 individuals over the age of 18 is randomly drawn from the National Registry of Norway. After receiving the data, persons over the age of

95 are excluded from the sample. From this subsample of the population, participants were invited by postal recruitment to participate in the study. Each data collection wave in the NCP is a survey of different questions, mostly related to policy attitudes in the population, but the platform also allows for economic games subject to approval from a scientific committee. At the time of planning the experiment in early 2015, we could not perform an informed power-analysis for the dictator game since we were unaware of studies that had tested for an effect of time constraints on dictator game giving.

Our data was collected as part of the fifth NCP wave conducted in 2015 (Skjervheim & Høgestøl, 2016a). We were the first to use this platform to study behavior in economic games, and we therefore expect subjects to be inexperienced with economic lab games. The arguments that have been raised against papers using Amazon Mechanical Turk to study intuitive prosociality – pointing out that the subject pool there is highly experienced (Rand et al., 2014; Rand, 2018) – should therefore not apply to our study.³

The experimental design is a 2 x 2 factorial design varying subjects' experience with a Dictator game and the time constraint they faced (time pressure vs. time delay). In total, 1390 subjects participated in the experiment. These participants were randomly drawn from the total Norwegian Citizen Panel sample to receive four treatments. Each of the participants received either a Time Pressure or a Time Delay condition. In addition, to examine the suggestion by Rand et al. (2014) that the time pressure effect is moderated by experience with economic games, we give about half the participants the treatments prior to exposure to a standard dictator game, and the other half obtained the treatments after exposure to a standard dictator game. This provides exogenous variation in subjects' experience with the game and allows us to causally study the effect of experience on the treatment effect of interest.

³ Reigstad, Strømmland & Tinghög (2017) report from several economic games studied in this panel. However, the Dictator game currently reported was conducted in 2015 and is the first economic game that was studied in the NCP. Prior to this, one economic experiment had been run, but involved a simple economic decision that did not involve other subjects (See Tjøtta, 2019).

In all experimental conditions we incentivized the decisions by informing participants that two people would be randomly drawn for real payment and subjects knew this when making their decision. Experimental evidence indicates that this random payment method produces similar decisions in the Dictator game as paying all subjects (Clot, Grolleau & Ibanez, 2018). All subjects decided how to divide 2000 NOK (223 USD) between themselves and another subject. Subjects could earn a maximum of 2000 NOK (223 USD) if they decided to keep the entire sum of money for themselves. We also incentivized treatment compliance by informing subjects that in order to be able to participate in the lottery they would need to decide within their allotted time.

In the **Time Delay** condition, after reading the experimental instructions concerning the Dictator game, participants entered a decision stage. All were told that in order to be able to be drawn out for real payment, they had to delay their decision for at least ten seconds. A timer counted upwards from zero and onwards in order to help subjects keep track of decision times. This design feature along with incentivizing compliance was implemented in order to facilitate compliance with the assigned time constraint.

In the **Time Pressure** condition, participants were instructed to decide *within* ten seconds in order to be able to receive payment. A timer counted down from ten seconds to keep track of the available time.

As mentioned above, half the subjects received treatment prior to a standard Dictator game, while the rest received treatment after a standard Dictator game. This game was structurally identical to the dictator game with time constraints. Table 1 summarizes the main features of the experimental design. The experimental instructions are presented in Appendix A.

Table 1. Main features of the experimental design

Time constraint	Experience with the Dictator game	
	No	Yes
Time Delay	N = 332	N = 389
Time Pressure	N = 324	N = 345

2.2. Testing for heterogeneous treatment effects

As there are several proposed sources of heterogeneity in the literature, a null finding in our main sample may be explained by other sources of heterogeneity unaccounted for in our analysis. According to the “Heterogeneity hypothesis”, the effect may be zero on average across study populations but systematically varies with observed and unobserved individual-specific variables (e.g. Rand, 2016). In contrast, if the null hypothesis is always true, there should be no systematic variation in the effects apart from sampling variation. While these two hypotheses are both able to explain a single null result, they have different predictions about the distribution of hypothesized conditional treatment effects.

Suppose independent research teams each test a simple null hypothesis, $H_0: \theta_k = 0$ against a two-sided alternative, where k indexes study population k . If there is heterogeneity in the effect, the θ_k will differ systematically across research teams apart from sampling variation. Suppose true effects θ_k depend on a vector of observable characteristics $\mathbf{x} \in X \subset \mathbb{R}^K$ so that $\theta_k = \mu(\mathbf{x}) + \vartheta_k$ where ϑ_k is a study-specific random effect with mean zero and variance τ^2 . Here we also assume that ϑ_k is mean independent of $\mathbf{x}$.

Denote by $\hat{\theta}_k$ the estimator used by the researcher. If the “Heterogeneity hypothesis” is true, then one could fail to reject the null hypothesis either because $E(\hat{\theta}_k) = 0$ or because the main test masks heterogeneity in relevant subgroups, so that the main test has low statistical power. By considering a treatment effect conditional on a set of theoretically relevant characteristics $\mathbf{x}^T$ one should expect $E(\hat{\theta}_k | \mathbf{x} = \mathbf{x}^T) > 0$, and if the researcher has correctly identified the characteristics that explain variation in the true effect, the average of all such conditional effects should be larger than zero and the power to detect statistically significant effects should be larger. The average power to detect an effect in the conditional distribution of effects is given by the average of the individual power estimates when averaging over all the theoretically relevant conditional effects. The expected fraction of significant results is $(1 - \beta)\pi + \alpha(1 - \pi)$, where π is the fraction of conditional effects that are true. For instance, if all effects are true and the statistical power is equal to the conventional 80% level, then 80% of the effects in the conditional distribution are expected to be statistically significant.

Now suppose there is no treatment effect heterogeneity so that $\tau^2 = 0$ and $\theta_k(\mathbf{x}) = 0 \forall \mathbf{x}$; the null hypothesis is true in any study population. In this case both the average unconditional effect and the average conditional effect (for any vector $\mathbf{x}^T$) is expected to be zero. Moreover, under the null hypothesis and a single homogenous effect the p-values follow a uniform distribution so that $F(p) = p$, and the expected fraction of statistically significant results at the 5% level is then equal to 5%.

To be able to interpret our findings clearly, we will consider the distribution of conditional treatment effects to single out whether we find evidence for genuine heterogeneity in the time pressure effect.

3. Results and discussion

3.1. Overall patterns

Our first result is that there are no discernable differences in giving in any of the four different experimental conditions.⁴ In fact, contrary to the prediction that intuition should promote prosocial behavior, the estimated time pressure effect for inexperienced subjects is actually negative, although small (-2 percentage points) and statistically insignificant in an OLS regression ($p = 0.145$, $N = 1368$).⁵ Importantly, this effect is rather precisely estimated; the 95 percent confidence interval ranges from -5 percentage points to 0.7 percentage points. We neither find an interaction effect among time pressure and experience (Coefficient = 0.024, $p = 0.215$), but the estimated effect is also here in the opposite direction than the one predicted by the SHH. Finally, experience has no causal effect on dictator game giving, although the estimated effect is in the negative direction of negative 1.9 percentage points but not statistically significant. The latter result contrasts with a recent paper by McAuliffe et al. (2018) who report that experience with a dictator game lowers subsequent sharing and cite this finding as support for the SHH.

To explain the lack of an effect of time constraints on giving, it is first necessary to assess whether the time constraint manipulation successfully affected decision times involved in the decision. Figure 1 displays the distribution of decision times by the time constraint the subjects were assigned to.

⁴ There are 1.6% missing observations, evenly distributed among the experimental conditions. Such a low rate of attrition is unlikely to influence the results. The fraction of missing observations in our experiment is also comparable to the overall rate of missing observations. For instance, on the central question in the survey of which party the participant would vote for if election tomorrow 1.1% responses are missing.

⁵ As we cannot a priori determine the direction of the expected effect, all hypotheses tests are two-sided.

Figure 1: Distribution of decision times by assigned time constraint (the conditions with the same time constraint are pooled)

Figure 1 indicates a clear distributional shift towards less time spent making the decision in the conditions with Time Pressure. In the Time Delay treatment, subjects on average spend 23 seconds making their decisions, while in the Time Pressure treatment, subjects on average spend 14.8 seconds, 8.5 seconds less than the subjects with Time Delay ($p < 0.001$, OLS-regression, table A.1.). The difference in decision times in the Time Pressure condition with and without prior exposure to a dictator game is not significantly different (coefficient = -0.77, $p = 0.436$), but the subjects with experience spend 3.5 seconds less time making their decision in the Time Delay condition ($p < 0.001$, OLS-regression, table A.1.).

Although our treatment successfully induced variation in time spent making the decision, a lot of subjects did not comply with the assigned time constraint. On average, 67.4 percent are compliers in the experiment, but the compliance rate is highly unevenly distributed

across the treatments; in the Time Pressure condition the compliance rate is only 28.7%, while the compliance rate in the Time Delay condition is 92%. Although this number qualitatively matches previous research in this literature (e.g. Bouwmeester et al., 2017; Rand, Greene & Nowak, 2012), a compliance rate of less than 30% is very low and may partially result from the representativity of our sample – typical experiments in this literature have used young and highly educated university students who may spend less time making their decision. In fact, the average compliance rate across the time pressure conditions for participants with at least college-level or university education is 46.8% and the compliance rate with time pressure for subjects below the age of 35 is 63.2%, indicating that the compliance rate for these subgroups is much higher and comparable to what is typically reported in other experiments using non-representative samples.

As our estimates capture an intention to treat effect, we also report results from an analysis with instrumental variables estimator. Although there is substantial non-compliance the treatment assignment significantly predicts actual treatment status ($t = 7.15$, $p < 0.001$). The two remaining assumptions is that the instrument (treatment assignment) should not impact on giving in the dictator game through other channels than the time spent making the decision (the *exclusion restriction*), and treatment assignment should influence the probability of being treated in the same direction for all subjects (the *monotonicity* assumption) (Angrist, Imbens & Rubin, 1996). Table A.2 summarizes the results from OLS and IV regressions. As the IV result is just a scaled version of the OLS result, for both specifications the estimated time pressure effect is negative for inexperienced subjects. Moreover, neither the OLS nor the IV result are statistically significant.

3.2. Testing the “Heterogeneity hypothesis”

There are several other proposed sources of heterogeneity in the literature on the cognitive foundations of prosocial behavior, and our overall result of no effect for both experienced and unexperienced subjects may mask effects of intuition in sub-populations. Rand et al. (2016) claim that the SHH predicts “intuitive altruism” only for women. Another potential explanation for our null result is that time constraints have opposing effects depending on subjects’ initial predisposition towards prosociality – some argue that intuitive decision making reinforces subjects’ initial predisposition (Alós-Ferrer & Garagnani, 2018; Chen & Krajbich, 2018). Under this interpretation, time pressure should make altruistic subjects more altruistic and work in the opposite way for selfish subjects. To test this prediction, we use the data from the unconstrained version of the Dictator game. We classify subjects based on their choice in the unconstrained dictator game and define “pro-social” subjects as those that share at least 50% of the endowment with the receiver.

Another potential source of variation in the treatment effect is the extent to which individuals trust their everyday interaction partners. It has been claimed that the SHH predicts that only individuals accustomed to cooperation paying off in their everyday lives will exhibit intuitively prosocial behavior – and that generalized survey measures of trust serve as a useful proxy for this measure (Rand & Kraft-Todd, 2014). The subjects in the Norwegian Citizen Panel survey also answered generalized trust questions and we split our sample in two based on their trust score. We define high trust subjects as those who gave a 5 or higher on a 10-point scale.

Finally, as the compliance rate was very low for the general sample but considerably higher among highly educated and young subjects, we will perform additional analyses where we restrict the analysis to these subgroups and the combination of these and other potential moderators. This also makes sense considering the recent discussion of the registered

replication of Rand et al. (2012) by Bouwmeester et al. (2017), who found a null effect that was largely explained by the original author as resulting from a low compliance rate (Rand, 2017).

In total the variables considered in our subgroup analysis yields $2^6 - 1 = 63$ different specifications for the distribution of conditional treatment effects. The results are displayed in figure 2.

Figure 2: Distribution of treatment effects in the distribution of conditional effects (A) and the corresponding distribution of p-values (B), along with kernel density estimate

The distribution of conditional effects in our data have a median of 0.006 and a mean of 0.012, while the mean p-value is 0.53. Out of the 63 p-values, 2 are significant at the 5% level, slightly less than the 3.15 expected from chance alone while 7 are significant at the 10% level, very close to the 6.3 expected from chance alone. Moreover, figure A.1. plots the p-values of our analysis against the quantiles of the uniform distribution and the data points are closely distributed along the straight line. We also carry out a one-sample Kolmogorov-Smirnov test based on the known theoretical reference distribution when the null hypothesis is true, $F(p) = p$. This test yields a p-value of 0.541. These results are unlikely to result from low power as the mean sample size in the subgroup analyses is equal to 288 and the median sample size is 193. However, if the underlying effect is true, we should expect the p-value to be decreasing in the

sample size whereas the p-value should be unrelated to the sample size if the underlying effect is zero (Hung et al., 1997). We find no evidence for a systematic correlation between the sample size and the p-value in our distribution of tests – in fact the estimated correlation is in the opposite direction than predicted when the alternative hypothesis is true (Spearman’s rho = 0.065, p = 0.613).

If there really was heterogeneity in the underlying effect, we should have expected a right-skewed distribution of p-values, the fraction of statistically significant tests should have been larger than 0.05 and the distribution of conditional treatment effects should have had a mean larger than zero. Moreover, the p-value should have been negatively related to the sample size. In light of our findings, the data seem more consistent with a model where there is no effect across study populations, and sampling variation alone can account for the results. The nonparametric estimate of the distribution of effects looks symmetric and tightly centered on a mean close to zero.

3.3. Interpreting our findings in light of previous literature

While we find evidence against the “Heterogeneity hypothesis” in our data, how can we explain the relatively large string of papers finding support for a heterogeneous effect? Following Ioannidis (2005), we consider the “post-study probability” (PSP) – the probability that a randomly drawn research finding in a given literature is true, conditional on statistical significance (Ioannidis, 2005),

$$PSP = \frac{(1 - \beta)\pi + \beta\pi u}{(1 - \beta)\pi + \beta\pi u + [\alpha + (1 - \alpha)u](1 - \pi)}$$

Where π is the prior, $(1 - \beta)$ is the statistical power, α is the significance level and u captures the extent to which research findings in this literature are “biased” through some form of

selective reporting. The PSP will be decreasing in u as long as $(1 - \beta) > \alpha$ (Ioannidis, 2005), which will typically hold. Following Maniatis, Tufano & List (2017) we can specify u as a function of the degrees of freedom available to researchers working in the literature ($d \in [0,1]$), the pressure to publish ($\xi \in [0,1]$) and editorial preferences for statistically significant findings ($\mu \in [0,1]$), $u = f(d, \mu, \xi)$, where it is assumed that $\nabla u > 0$. If μ and ξ are non-zero, an increase in the degrees of freedom available to the researchers working in the literature will be associated with a decrease in the PSP. A research literature where the underlying hypothesis may be corroborated only by appealing to “main” tests and not subgroup analyses will, *ceteris paribus*, have a lot less degrees of freedom than a research literature where the underlying hypothesis may be supported by appealing to statistically significant subgroup analyses. For instance, in our heterogeneity analysis alone we reported 63 different tests that could all be rationalized by reference to previous literature; none of which could be used in support of the “intuitive prosociality” hypothesis if the hypothesis strictly predicted a positive effect in the main sample.

4. Concluding remarks

Some suggest that there is an effect of intuitive decision processes on prosocial decisions, but that the effect is heterogeneous. We study causally the effect of the most important proposed source of heterogeneity – subject experience – and consider the distribution of conditional treatment effects where we simultaneously test for several previously proposed sources of heterogeneity in this literature. We find no clear indication of heterogeneous effects – rather, the distribution of effects has a mean very close to zero and only 2 of the tests pass the conventional threshold associated with statistical significance, slightly less than the number expected from chance alone. This evidence is consistent with the hypothesis of a homogenous zero effect of time pressure on prosocial behavior, and that the flexibility of the research

hypothesis combined with researcher degrees of freedom may have caused the appearance of heterogeneity in this literature.

An alternative explanation of our findings is that factors unaccounted for in our analysis explain our lack of support for the “intuitive prosociality” hypothesis. This could occur if our sample or experimental design in some features systematically differ from typical experiments previously reported in the literature. In terms of the model of heterogeneous effects outlined in section 2, $\vartheta_k < 0$ – the error term is a negative draw from the symmetric distribution of study effects, canceling out against a positive mean effect in the distribution of conditional effects that would have appeared with a draw of ϑ_k closer to zero. While this objection cannot be ruled out, a recent meta-analysis by Fromell, Nosenzo & Owens (2018) of the effect of intuition on altruism considers several other sources of heterogeneity than the ones considered in our paper, such as the type of manipulation used to study intuitive pro-sociality, whether students were used, which country the experiment was run in, the stakes involved in the decision and which framing was used. They too find very little heterogeneity and find no evidence that any of these factors influence the effect size. Another possibility is that the “Heterogeneity hypothesis” is true, but that the factors we have considered are not the correct ones. An interesting feature of the latter criticism is that it would also imply that previous papers have identified the wrong variables.

In summary, there seems to be little evidence for an effect of cognitive processing manipulations on prosocial behavior in the Dictator game. Altogether, these findings are good news for researchers interested in modeling social preferences – it seems that we do not need to explicitly model the cognitive system primarily involved in making the decision. We also believe that our analysis approach should be more broadly adopted in Experimental Economics. There is a concern about the replicability of experimental findings in Economics (Camerer et al., 2016), and the distribution of p-values in lab experiments in economics seem to suggest the

presence of selective reporting of statistically significant results (Strømland, 2019). The analysis approach adopted in this paper makes it very hard to present a single result as evidence for an underlying theory of interest as it is very hard for a researcher to “manipulate” an entire distribution of effects.

5. References

- Alós-Ferrer, C., & Garagnani, M. (2018). The cognitive foundations of cooperation. *University of Zurich, Department of Economics, Working Paper*, (303).
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434), 444-455.
- Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, 113(4), 936-941.
- Bouwmeester, S., Verkoeijen, P. P., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., ... & Evans, A. M. (2017). Registered replication report: rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, 12(3), 527-542.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637.
- Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7), 1131-1139.
- Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, 9(1), 3557.
- Clot, S., Grolleau, G., & Ibanez, L. (2018). Shall we pay all? An experimental test of Random Incentivized Systems. *Journal of Behavioral and Experimental Economics*, 73, 93-98.
- Døssing, F., Piovesan, M., & Wengström, E. (2017). Cognitive load and cooperation. *Review of Behavioral Economics*, 4(1), 69-81.
- Everett, J. A., Ingbreten, Z., Cushman, F., & Cikara, M. (2017). Deliberation erodes cooperative behavior—Even towards competitive out-groups, even when using a control condition, and even when eliminating selection bias. *Journal of Experimental Social Psychology*, 73, 76-81.
- Fromell, H., Nosenzo, D., & Owens, T. (2018). *Altruism, Fast and Slow? Evidence from a Meta-Analysis and a New Experiment*. Retrieved from: <https://ideas.repec.org/p/not/notcdx/2018-13.html>

- Goeschl, T., & Lohse, J. (2018). Cooperation in public good games. Calculated or confused? *European Economic Review*, 107, 185-203.
- Hauge, K. E., Brekke, K. A., Johansson, L. O., Johansson-Stenman, O., & Svedsäter, H. (2016). Keeping others in our mind or in our heart? Distribution games under cognitive load. *Experimental Economics*, 19(3), 562-576.
- Hung, H. J., O'Neill, R. T., Bauer, P., & Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 11-22.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Leamer, E.E. (1983). Let's Take the Con Out of Econometrics. *American Economic Review* 73(1): 31-43.
- Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, 104(1), 277-90.
- Maniadis, Z., Tufano, F., & List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal* 127 (605), F209-F235.
- McAuliffe, W. H., Forster, D. E., Pedersen, E. J., & McCullough, M. E. (2018). Experience with anonymous interactions reduces intuitive cooperation. *Nature Human Behaviour*, 2(12), 909.
- Merkel, A. L., & Lohse, J. (2018). Is fairness intuitive? An experiment accounting for subjective utility differences under time pressure. *Experimental Economics*, 1-27.
- Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3), 631-647.
- Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, 27(9), 1192-1206.
- Rand, D. G. (2017). Reflections on the time-pressure cooperation registered replication report. *Perspectives on Psychological Science*, 12(3), 543-547.
- Rand, D. G. (2018). Non-naïvety may reduce the effect of intuition manipulations. *Nature Human Behaviour*, 2(9), 602.
- Rand, D. G., Brescoll, V. L., Everett, J. A., Capraro, V., & Barcelo, H. (2016). Social heuristics and social roles: Intuition favors altruism for women but not for men. *Journal of Experimental Psychology: General*, 145(4), 389.
- Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427-430.

- Rand, D. G., & Kraft-Todd, G. T. (2014). Reflection does not undermine self-interested prosociality. *Frontiers in Behavioral Neuroscience*, 8, 300.
- 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.
- Reigstad, A. G., Strømland, E. A., & Tinghög, G. (2017). Extending the cooperative phenotype: Assessing the stability of cooperation across countries. *Frontiers in Psychology*, 8, 1990.
- Sala-i-Martin, X. (1997). I just ran two million regressions. *American Economic Review* 87(2):178-183.
- Schulz, J. F., Fischbacher, U., Thoni, C. & Utikal, V. (2014) Affect and fairness: dictator games under cognitive load. *J. Econ. Psychol.* doi:10.1016/j.joep.2012.08.007.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2), 534.
- Skjervheim and Høgestøl. 2016a. Norwegian Citizen Panel, Fifth Wave Methodology report. Technical report Ideas2Evidence
- Strømland, E. (2019). Preregistration and reproducibility. *Journal of Economic Psychology* (In press).
- Strømland, E., Tjøtta, S., & Torsvik, G. (2016). Cooperating, fast and slow: Testing the social heuristics hypothesis. *CESifo Working Paper Series No. 5875*. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2780877
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702-712.
- Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., ... & Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498(7452), E1.
- Tinghög, G., Andersson, D., Bonn, C., Johannesson, M., Kirchler, M., Koppel, L., & Västfjäll, D. (2016). Intuition and moral decision-making—the effect of time pressure and cognitive load on moral judgment and altruistic behavior. *PloS one*, 11(10), e0164012.
- Tjøtta, S. (2019). More or Less Money? An Experimental Study on Receiving Money. Fortcoming, *Journal of Behavioral and Experimental Economics*.
- Verkoeijen, P. P., & Bouwmeester, S. (2014). Does intuition cause cooperation? *PloS one*, 9(5), e96654.

Appendix

Appendix A: Experimental Instructions (Translated from Norwegian)

Instructions (common to both time constraint treatments)

As a participant in the Norwegian Citizen Panel, you can be drawn for an extra monetary prize. 2 people are drawn for payment: Person A and person B. Person A receives 2000 NOK and will choose how much to share with person B. Go to the next page in order to make a decision.

Decision stage, Time Delay

In order to participate in the lottery, we ask you to delay your decision for at least 10 seconds.

If I receive 2000 NOK, I choose to give the following amount to the other person (between 0 and 2000)

[Slider from 0 to 2000 NOK]

Decision stage, Time Pressure

In order to participate in the lottery, we ask you to answer within ten seconds.

If I receive 2000 NOK, I choose to give the following amount to the other person (between 0 and 2000)

[Slider from 0 to 2000 NOK]

Instructions for the unconstrained dictator game

As a participant in the Norwegian Citizen Panel, you can be drawn for an extra monetary prize.

2 people are drawn for payment: Person A and person B. Person A receives 2000 NOK and will choose how much to share with person B.

If I receive 2000 NOK, I choose to give the following amount to the other person (between 0 and 2000)

[Slider from 0 to 2000 NOK]

Appendix B: Regression tables and figures

Table A.1: OLS estimates of the treatment effects on decision times

	(1)
Time Pressure	-8.522*** (0.797)
Experience	-3.500*** (0.842)
Time Pressure*Experience	-0.770 (0.987)
Constant	23.32*** (0.665)
<i>N</i>	1368
Robust standard errors in parentheses * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$	

Table A.2: OLS and IV estimates of the Time Pressure effect

	OLS	IV
Time Pressure	-0.0212 (0.0145)	- 0.10 (0.0707)
Experience	-0.0193 (0.0129)	-0.0280 (0.0182)
Time Pressure*Experience	0.0236 (0.0190)	0.106 (0.0755)
Constant	0.496*** (0.00968)	0.504*** (0.0142)
<i>N</i>	1368	1368

Robust standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A.1: Plot of the p-values in our heterogeneity analysis against the quantiles of the uniform distribution