A New Mechanism to Alleviate the Crises of Confi dence in Science With An Application to the Public Goods Game

Luigi Butera, Philip J Grossman, Daniel Houser, John A List, Marie Claire Villeval

A New Mechanism to Alleviate the Crises of Confidence in Science - With An Application to the Public Goods Game*

Luigi Butera[†], Philip J. Grossman[‡], Daniel Houser[§], John A. List[¶] and Marie Claire Villeval[‖]

June 17, 2022

Abstract

Recently a credibility crisis has taken hold across the social sciences, arguing that a component of Fischer (1935)'s tripod has not been fully embraced: replication. The importance of replications is not debatable scientifically, but researchers and journals' incentives are not sufficient to encourage replications. We analyze a novel, decentralized approach promoting replications through beneficial gains between scholars and editors. We highlight the trade-offs involved in seeking independent replications before submission to journals, and demonstrate the operation of this method via an investigation of the effects of Knightian uncertainty on cooperation in public goods games, a pervasive feature but largely unexplored in the literature.

Keywords: Replication, science, public goods, uncertainty, experiment.
JEL codes: A11, C18, C92, D82.

* *Acknowledgements*: For useful comments we are grateful to Tommy Andersson, Alec Brandon, Gary Charness, Lucas Coffman, Giacomo De Giorgi, Salvatore Di Falco, Anna Dreber, Lata Gangadharan, Håkan Holm, David Jimenez-Gomez, Olof Johansson Stenman, Johanna Mollerstrom, Fatemeh Momeni, John Nye, Giovanni Ponti, Adam Sanjurjo, Jurre Thiel, Joe Vecci, Roberto Weber, as well as seminar participants at University of Chicago, Lund University, University of Alicante, George Mason University. On Grossman's side, this research has been funded by the Australian Research Council (DP130101695). On Villeval's side, this research has benefited from the support of IDEXLYON from Université de Lyon (INDEPTH) within the Programme Investissements dAvenir (ANR-16-IDEX-0005), and of the LABEX CORTEX (ANR-11-LABX-0042) within the program Investissements d'Avenir (ANR-11-IDEX-007) operated by the French National Research Agency (ANR). We thank Ariel Listo for outstanding research assistance. The study was pre-registered at the AEA RCT registry: AEARCTR-0002142.

[†]Copenhagen Business School. lbu.eco@cbs.dk. Corresponding author
[‡]Monash University. philip.grossman@monash.edu.
[§]George Mason University. dhouser@gmu.edu.
[¶]University of Chicago and NBER.jlist@uchicago.edu
[‖]Univ Lyon, CNRS, GATE UMR 5824, F-69130 Ecully, France; IZA, Bonn; villeval@gate.cnrs.fr.

1 Introduction

Economists, much like scientists, astronomers, and meteorologists, have traditionally relied on observational data to understand the world. While each of these empirical enterprises differs in subject matter, they share a common property: they all rely on important natural disturbing influences to settle differences. While this empirical approach remains an important intellectual pursuit in economics, one recent trend has been to take a less passive approach to empirical work, and generate exogenous variation through laboratory and field experiments (Harrison and List, 2004). Yet, this expansion presents concomitant challenges. How can we ensure that knowledge generation evolves in an optimal manner? How can markets and market forces be used to ensure that this happens within economics?

We address these questions by focusing on one aspect of the experimental approach: replication. Replicating empirical studies, particularly those whose findings are at odds with the current state of knowledge, can significantly accelerate the advancement of economic science. This is crucial to assess the validity of novel findings, and to quantify the uncertainty about their effect sizes (DellaVigna and Linos, 2022). Further, just a few successful replications can robustly increase the import of new findings, improving their placement and visibility in academic journals and civil society (Maniadis et al., 2014).

Nonetheless, the field of economics, in its current state, presents few strong incentives to replicate, both for researchers and academic journals. Once a study has been published, researchers have little incentive to replicate their own findings, given the high premium for novelty in economics.[1] Similarly, successfully replicating someone else's work does not grant high returns as in other disciplines like medical sciences, and failing to replicate someone's study, for instance by finding a null result, often results in the study not being published (Andrews and Kasy, 2019), or not even written up (Franco et al., 2014).[2]

Academic journals, while interested in publishing robust research, face important tradeoffs in promoting replications from the top. First, imposing ubiquitously tighter replication obligations might disproportionately hurt scholars with limited resources, junior scholars, and minorities, as they might not have the financial bandwidth to replicate their own work prior to publication. Moreover, journals must compete to attract and quickly publish

[1] Replication of researchers' own work before publication is also rare. See Kessler and Meier (2014) for an example.

[2] On the other hand, finding opposite, significant results can generate enmity, further reducing the incentives to replicate (Coffman and Niederle, 2015).

innovative results, thus creating a trade-off between robustness and novelty.[3]

Because novel results attract attention, and are generally sought after by academic journals much more than replication studies, an important question of incentives arises.[4] Yet, such lack of incentives to replicate is problematic, as new findings might be at odd with existing knowledge. Such novel studies may be false positives simply due to the mechanics of statistical inference (Maniadis et al., 2014; Coffman and Niederle, 2015; Dreber et al., 2015; Coffman et al., 2017). Similarly, new and surprising studies may suffer from low power or weak initial support, and thus may be dismissed even though they point toward an economic association that is ultimately true.[5] Uncertainty about the reliability of experimental studies may in part explain the documented decline in the number of experimental papers – both lab and field – published in top-tier journals over the last two decades (see Nikiforakis and Slonim (2019) and Fréchette et al. (2021)).

This paper analyzes and demonstrates the application of a novel and simple replication mechanism that generates mutually beneficial gains from trade among the authors of a novel study, other scholars working in the same area of research, and editors. In this mechanism we analyze, the original investigators, upon completing an initial study, write a working paper version of their research. While they can share their working paper online, thus establishing the paternity of the idea, they commit to not submitting the work to a journal for publication. They instead invite other researchers to coauthor and publish a second, yet-to-be-written paper, provided that researchers are willing to replicate independently the experimental protocol in their own research facilities. Once the team is established, but before the replications begin, the replication protocol is preregistered and

[3]Further, blanket requirements for replications would likely be inefficient, as many new studies that are coherently embedded in a well established literature may not necessarily warrant replications as a necessary condition for publication (Maniadis et al., 2017).

[4]While this paper focuses on the close replication of an existing experimental design, other types of replications are relatively rare as well, such as obtaining published datasets to replicate the results, or investigating a research question using a different design and setting (see e.g. Hamermesh (2007)). This contrasts with an increasing demand for robustness in economic research (see, for example, Benjamin et al. (2018); Camerer et al. (2018)

[5]For example, the paper of Fischbacher and Föllmi-Heusi (2013) took several years to get published, although the paradigm of the die-under-the-cup has become extremely influential and used in more than 90 studies since 2013 (see the meta-analysis of Abeler et al. (2019)). Vernon Smith reports that there was a false prevailing belief that transparency in asset values would prevent price bubbles in the early eighties; thus, initially, no one believed the results of his famous experiment with Suchanek and Williams, in which they found that values in use conflict with values in exchange (Smith et al., 1988). It was considered "an Arizona phenomenon." The first asset paper has eventually been published in *Econometrica*, but after three years of revisions and mostly negative reviews. According to V. Smith, the reason this research became popular is that the results were replicated by others (Smith, 2018).

referenced back in the first working paper.[6] This guarantees that all replications, both successful and unsuccessful, are properly accounted for. The team of researchers then writes the second paper, which includes all replications, and submits to an academic journal.

The mechanism we analyze is a decentralized "price-driven" approach that alleviates the incentive problem that researchers and editors face. The choice of replicating involves some costs, and we highlight in a model the conditions under which this mechanism might be chosen over the *status quo* publishing approach. In our model, researchers can improve completed research papers through a combination of labor-intensive and capital-intensive investments: Adding robustness checks and alternative models, hiring more research assistants, purchasing additional data or – as in our case – by replicating. Because any investment is time consuming, including replications, this reduces the incentives to invest for researchers who face short and long-run time constraints. However, we show that long and short-run impatience, which is likely experienced by nontenured researchers, can delay – not accelerate – the publication process: Impatient researchers forego time-consuming investments today, despite the fact that these investments can improve the quality and placement of their research, but remain too selective in the choice of target journals. A first condition for the mechanism to be incentive-compatible is that time-constrained researchers are sophisticated about their own time preferences, since the mechanism operates as a binding commitment device that improves overall utility. A second condition is that journal editors prefer empirical results that have been independently replicated, *ceteris paribus*.

As we explain in our theory section, the ultimate incentive of the mechanism for both original authors and coauthors lies in the possibility of generating more robust research, which could consequently lead to better publications that would otherwise be hard to attain. As detailed, there are many other approaches that could promote more replications, many of which require coordination among journals. The mechanism we describe should instead be intended as a voluntary mechanism that responds to a practical problem: How can individual scholars increase the robustness of their results – and consequently the positioning of their research – at a relatively low cost?

Note that the approach we analyze in this paper is different from simply collecting more data in the first place. While the latter helps to increase power and the ability to detect smaller effect sizes, the mechanism under investigation is instead geared towards

[6] In particular, in our pre-registration we included the original manuscript, and consequently the whole original analysis.

generating stronger Bayesian posteriors, as well as accounting for potentially important differences across experimental environments. As we discuss in section 2, in terms of inference a single large study may deliver lower posterior beliefs compared to a collection of smaller, lower-powered successful replications. Intuitively, this is because it is harder to consistently find a significant result in a series of relatively smaller-scale replications than finding the same result in a single better powered study (see Maniadis et al. (2017)).

The mechanism we investigate applies generally to any empirical research, but in this paper we illustrate how it can be used for experimental research. Precisely, we test its applications to one of the most active areas of research in experimental economics: public goods games (see Ledyard (1995); Chaudhuri (2011); Villeval (2020) for reviews). Within a public goods game setting, we investigate how the presence of Knightian uncertainty (ambiguity) over the quality of the public good affects cooperation rates. The question is important since returns from public goods and social programs in real settings are, more often than not, intrinsically uncertain and difficult to quantify *ex-ante*.

Quite surprisingly, the original, voluntarily unpublished investigation (Butera and List, 2017) found that Knightian uncertainty facilitates cooperation, thereby reducing the decay of cooperation over time typically observed in standard public goods games. Following our replication mechanism, the working paper was distributed online, but voluntarily never submitted to a journal for publication. The current paper reports results from the original experiment, conducted at Georgia State University, and three follow-up replication studies carried out at GATE-Lab in Lyon, France, at the ICES lab at George Mason University, United States, and at MonLEE Lab at Monash University, Australia.

We find evidence in two out of three replications that Knightian uncertainty positively affects cooperation when the quality of the public good is low.[7] Yet, when considering the basic result of whether Knightian uncertainty facilitates overall cooperation, the original results do not replicate using a conservative replication test. We take this key insight and explore the inference one takes from a Bayesian analysis of the Post-Study Probability. In short, we find that while inference critically depends on the nature of priors, with surprising results such as ours, the independent replications allow us to rule out the idea that Knightian uncertainty plays an economically significant role in cooperative decisions.

One can imagine that if we had taken the traditional approach of discovery and publication, followed by "fighting about the results that do not replicate the original insights"

[7] We find similar results in the same direction at $p < 0.05$ for two-sided tests.

later in journals, the time and resources used to reach this conclusion would have been many times greater than those expended in this case. In this manner, our study represents a first attempt at implementing a new replication mechanism that has many attractive features.

Beyond its methodological contribution, our paper contributes to several strands in the literature. First, it contributes to the small, but growing, literature on mechanism design for replications. Three main approaches have been proposed in the literature: a top-down institutional approach, a bottom-up cultural approach, and a market approach. A top-down institutional approach requires the involvement of professional organizations, funding agencies, and academic journals in promoting a culture of replication. One possibility is the creation of academic journals that openly invite submission of replications.[8] This approach, while desirable, might not fully address the fact that replication studies generally carry low returns in terms of academic prestige, and this in turn may hinder the popularity of new replication journals (Maniadis et al., 2015). Another possibility is proposed by Coffman et al. (2017), who suggest that premier journals include a simple one-page "replication reports" section. Major journals could indeed promote more replications by leveraging the prestige of their outlets, but in practice such an attempt has not been made yet.[9] In practice, academic journals face several challenges in coordinating and promoting replications from the top.

A second type of solution is a bottom-up, cultural approach aimed at changing social norms within the academic community regarding replications. For instance, Coffman et al. (2017) propose the norm of citing replication work alongside the original, granted of course

[8]For instance, *Experimental Economics* – as well as its companion journal *Journal of the Economic Science Association*– clearly state in their aims and scope statute to focus on publishing "[..] *article types that are important yet underrepresented in the experimental literature (i.e., replications, minor extensions, robustness checks, meta-analyses, and good experimental designs even if obtaining null results)*".

[9]While the allure of publishing in top journals may encourage scholars to produce and publish replication studies, Hamermesh (2017) points out that the opportunity cost of devoting space to replications that arguably do not generate the same interest in readership as original articles (Whaples, 2006) might be too high. In the same article, regarding nonexperimental papers, Hamermesh suggests that major journals could, in principle, recruit a cadre of replicators to verify an accepted article. However, he points out that there would be very little incentives for scholars to become replicators, and there would still be the question of who "guards the guardians". One attempt of this approach has been taken by Drazen et al. (2019), who tested a proof-of-concept method in which a journal – in their case Journal of Public Economics – contracts for a replication between acceptance and publication of the paper. In their case, the journal invited the authors of several accepted papers to voluntarily opt-in this mechanism, with guarantee that the replication outcome would not alter the acceptance decision. Their article reports on one replication of the study by Drazen and Ozbay (2019), who accepted to join this exercise.

that the replication effort is ultimately published.[10] From that perspective, creating repositories listing all papers that have been replicated citing both the original studies and the papers replicating them, would facilitate the transition to a new culture of replication. For example, the Economic Science Association has initiated a replication log for the articles published in its journals. Further, Maniadis et al. (2015) suggest that using the number of replications of one's experimental work (both successful and failed) as a metric for one's research quality (*e.g.*, for funding and promotion purposes) might help reduce the enmity among researchers that replication often induces.

A third solution, which is closer to the approach explored in this paper, is a decentralized market approach to replications. In seminal work, Dreber et al. (2015) explore the replicability of recent publications in top psychology journals by using prediction markets populated by graduate students and professors. In each market, participants trade contracts that pay real incentives if the study is replicated. Dreber et al. (2015) find that market prices are strongly correlated with the success of replications. We view the mechanism explored in this paper as a complement to this approach, in that it leverages prices (in the case of our paper, in the form of willingness to pay the costs of replications), but does not require an external party to coordinate replications (see also Landy et al. (2020)). Another noticeable advantage of our mechanism is that publishing an original paper with its own replications combines the interest of the readership for the original research with that of the robustness provided by the replications. Furthermore, the mechanism we analyze is particularly well-suited to handle studies whose surprising results are very likely to be met by low priors (see also Camerer et al. (2016, 2018)).

Our study also contributes to a second literature that is the debate on the scientific value of null results. Insignificant results are notoriously difficult to publish (Ziliak and McCloskey, 2008), and the notion that such results are noninformative is common among economists (Abadie, 2020). Andrews and Kasy (2017) estimate the probability of publishing significant results being 30 times higher than publishing null results. Not only are significant results more likely to publish well, they are also more likely to be written up in the first place (Franco et al., 2014). Yet, insignificant results also provide important information (Kessler and Meier, 2014; Abadie, 2020), particularly in contexts where there is no *a priori* reason to believe in a zero effect of an intervention (*e.g.*, Abdulkadiroglu et al. (2014); Cesarini et al. (2016); de Ree et al. (2018); Meghir et al. (2018)). While it cannot

[10] For estimates on the rate replications in leading journals, see Berry et al. (2017).

be directly used to confirm a null hypothesis, the Post-Study Probability derived from a series of independent failed replications provides critical information about an economic phenomenon. Such a simple approach allows scholars to maintain a frequentist approach to economic analysis, while providing scholars with a Bayesian toolkit to assess whether they should be more or less likely to reject the null hypothesis. This should increase the scientific value of studies combining independent draws of null results, and therefore increase the academic returns from completing and publishing such studies.

Finally, our paper contributes to the literature on the private provision of public goods, one of the most studied decision-making environments in the field of experimental economics (e.g. Andreoni (1995)). A feature common to these studies is the absence of uncertainty about the value of the public good. Only a handful of papers depart from certainty about the value of the public good (see, e.g., Fisher et al. (1995), Levati et al. (2009), Gangadharan and Nemes (2009) and Theroude and Zylbersztejn (2020). Our work departs from these studies in several key ways. First, none of these studies directly addresses the question of how social dilemmas are affected by irreducible ambiguity. One notable exception, conducted concurrently to our original study, is Bjork et al. (2016) who also allow the marginal return to contributions to be ambiguous. Interestingly, as in our replications, they find that uncertainty does not have a significant impact on the inclination to cooperate. Second, our parameter space allows us to investigate the effect of Knightian uncertainty over a rich set of situations, from social dilemmas to situations where it might be socially optimal not to fund the public good, and to cases where fully contributing might be a Nash equilibrium. Third, our design privately provides subjects with noisy signals, similar to the common value auction literature (see Harrison and List (2008)). This structure allows us to capture a critical feature of real-life public goods: When choosing whether and how much to contribute, individuals must take into account that other contributors, like themselves, may hold optimistic or pessimistic beliefs about the value of the public good.

The remainder of this paper is organized as follows. The next section presents our incentive-compatible mechanism for replication. Section 3 develops a theoretical framework of the replication dilemma. Section 4 introduces our experimental design and procedures. Section 5 presents our experimental results. Finally, section 6 discusses the limitations and challenges of our mechanism, and concludes.

2 A Simple Incentive-Compatible Mechanism for Replication in Economics

This section builds the intuition for the importance of replicating, then details the replication mechanism proposed by Butera and List (2017) (BL, hereafter). Finally, it shows why the mechanism under investigation may be particularly well-suited to original studies that are likely to be met by low priors.

2.1 The Benefits of Replications

Maniadis et al. (2014) propose a simple Bayesian framework to evaluate how novel results should move researchers' priors.[11] Let us start with the simplest case of updating after observing results from one study. Let π be the prior that a given scholar has about a given scientific relationship. Call α the significance level of an experiment investigating such relationship, and $(1-\beta)$ the power of the experiment. The Post-Study Probability (PSP, hereafter) that a given scientific association is true can be computed using the following formula:

$$\text{Post-Study Probability} = \frac{(1-\beta)\cdot\pi}{(1-\beta)\cdot\pi + \alpha(1-\pi)} \tag{1}$$

where $(1-\beta)\cdot\pi$ represents the probability that a true result is declared true for any given prior π, and the denominator represents the probability that *any* result is declared true (*e.g.*, $\alpha(1-\pi)$ is the probability of a type I error given prior π). So, for instance, if a scholar believed that a given, not-yet-established scientific result had a 1% chance of being true at $\alpha = 5\%$ level and power $(1-\beta) = 80\%$, after observing one study establishing that result, he would update his priors to 13.9%. Another scholar holding priors of 10% would update the post-study probability to 64%.

Now suppose that subsequent replications of the study fail to find an effect at the same $\alpha = 5\%$. Figure 1 shows how Bayesian scholars holding initial priors $\pi = 1\%$ and 10% should update their posteriors based on subsequent failed replications. It can be easily seen that with three or more failed replications, the post-study probability converges toward zero, regardless of the initial priors. Conversely, as shown in figure 2 just a few successful replications allow convergence of PSP, even for low initial priors.

[11] Their approach builds on a formal methodology developed in the health sciences literature (Wacholder et al., 2004; Ioannidis, 2005; Moonesinghe et al., 2007).

Successful replications increase posteriors. Yet, equation 1 shows that the Post-Study Probability also increases with the power of the experiment, $1-\beta$. Thus, an important question is how inference improves by simply increasing the size of the experiment (i.e. increasing power), as opposed to conducting time-consuming replications.

If all observations are i.i.d., then from an econometric perspective the results from a single large study would be equivalent to a collection of small-scale replications, granted that they have the same total sample size. However, they may not be in terms of inference.

Consider again the case of low priors $\pi = 1\%$ and a well powered single study (e.g. $1-\beta = 80\%$). How does the PSP from this well-powered study compare to the PSP from a collection of replications of a lower powered study?

To calculate that, we can compare the PSP from a single high powered study to the PSP after observing a varying number of lower-powered successful replications i out of a total of n replication attempts.[12] Figure 3 fixes priors to $\pi = 1\%$ and reports results from this exercise for replication studies with power $1-\beta = 0.2$, and $\alpha = 0.05$ under two scenarios: The total number of studies is $n = 5$, or $n = 10$. The x-axis reports the number of successful replications, where 0 means that only the original study is significant. The y-axis reports the corresponding PSP. For comparison, Figure 3 also plots the horizontal PSP for priors $\pi = 1\%$ from a single, high powered study (e.g. $1-\beta = 0.8$) under two scenarios: First, for results significant at $\alpha = 5\%$, which would result in a $PSP = 13.9\%$. Second, for results significant at $\alpha = 1\%$, which instead would move priors $\pi = 1\%$ to $PSP = 44.7\%$. It's easy to see that even with only one-fourth of the power, just three successful replications (out of either five or ten total studies) generate higher posteriors than a single study with power equal to 80% finding a significant result at either 5% or 1% level.

The message from these illustrations is clear: First, a few independent replications allow for a wide range of beliefs to converge, regardless of initial priors. Second, replications are a highly effective way to increase posteriors. We suspect that this is why Fischer's (1935) original tripod included replications as one key feature besides randomization and blocking. As mentioned, however, there are few professional incentives for a wider and

[12] To compute the PSP for replication studies we can note that, conditional on the result being true, the probability of obtaining i successful replications out of a total number of n studies, follows a binomial distribution with probability of success equal to $1-\beta$, so $bin(1-\beta, i, n)$. Similarly, conditionally on the result being false, the probability of i false positives in n replication studies follows a binomial distribution $bin(\alpha, i, n)$ (Maniadis et al., 2017). Thus, the PSP from a collection of replications can be calculated as follows: Post-Study Probability$^{Rep} = \frac{bin(1-\beta,i,n)\cdot\pi}{bin(1-\beta,i,n)\cdot\pi + bin(\alpha,i,n)\cdot(1-\pi)}$

more systematic use of replications. As such, a simple incentive-compatible mechanism to promote replications can be useful.

2.2 The Replication Mechanism

We analyze a simple mechanism based on the notion of mutual gains from trade between the original authors of a novel study and other scholars interested in the same research topic. The mechanism we investigate is detailed for experimentation, but could easily be adapted to more general empirical exercises. The approach follows four steps.

Step 1: Upon completion of data collection and analysis of a new experiment, the original authors find a significant result. They commit to writing a working paper using the original data, but agree that they not submit it to a refereed journal. After calculating the minimum number of replications necessary to substantiate their results given their design, the original authors offer co-authorship of a second paper to other scholars who are willing to replicate independently the exact experimental protocol at their own institution, using their own financial resources. We believe that the first step is to establish the robustness of the initial idea. This is different from conducting additional treatments, which is what is expected from research meant to be published independently. This is why the mechanism analyzed here proposes exact replications. For field experiments, exact replications might be difficult to achieve. However, a viable solution would be to retain the original design, but test it in different environments. This would maintain the focus on the original question, and also allow to assess its validity in relation to the Selection, Attrition, Naturalness, and Scaling (SANS) framework proposed by List (2020).[13]

There is a mutual understanding that the second paper is the only paper that will be submitted to refereed journals upon completion of all replications, and that it will include an analysis of the original dataset and all replication datasets. There is also a mutual understanding that the second paper will reference the first working paper, and that the latter will be coauthored only by the original investigators. The reference to the first working paper serves a dual purpose: it enables the original authors to credibly signal the paternity of the original research idea (a feature that might be particularly desirable for early career scholars) and, as explained below, it provides a binding commitment device

[13]For example, subsequent replications could test the original idea in a context with a different level of selection into the experiment, attrition, naturalness, or a different environment that could speak to the scalability of the results.

for original authors and other scholars alike that increases the credibility and feasibility of the replication strategy.

Step 2: Once an agreement has been reached with scholars willing to replicate the original study, the original authors preregister the replication protocol with the American Economic Association RCT registry, or any other open access research registry. The registered protocol includes details about the experimental protocol and materials (*e.g.*, the instructions) and the data analysis and findings of the original study.[14] It lists the names and affiliations of the scholars who will replicate the study, and provides a tentative timeline for replications. All parties agree that only the replications listed in the AEA preregistration will be included in the second paper.[15]

Step 3: Once step 2 is completed, the original authors include in the first working paper a section describing the replication protocol, the list of scholars who will replicate, and the reference number for the AEA preregistration. The original authors then post (or update) their first working paper online.

Step 4: Replications are conducted, data is collected, and the second working paper is written and submitted to a refereed journal by the original authors and the other participating scholars.

2.3 New Incentives for Replications

While the mechanism we investigate provides direct incentives for scholars to replicate different kinds of empirical studies, we believe that it is best suited for studies that are likely to suffer from low priors, and to be particularly beneficial to researchers at the early stages of their careers. There are two main reasons for this.

First, as shown in Section 2.1, small deviations in priors yield large changes in posteriors when priors are low – for instance, $\pi < 50\%$. As a result, the journal placement of a novel study may critically depend on relatively small differences in referees' priors. Because an

[14] We did include the original working paper in the preregistration, therefore including the actual analysis.

[15] The reason for listing the replications and the replications team in the preregistration is twofold: First, it provides a commitment device for all scholars involved in the project. Second, and most importantly, it provides a credible signal about the total number of replications that will be conducted. This is critical to avoid file drawer problems.

article cannot be submitted to the same journal twice, scholars incur the risk of underplacing their work in terms of academic publishing, even when the research is technically sound and substantially interesting. By replicating their research, scholars can increase the probability of successful publication in higher quality journals. If the replication is successful, then the priors of referees and editors would not impact the paper's reception. This is because successful replications induce referees' posterior beliefs to converge, regardless of their priors.[16] For any given journal, a successfully replicated study would therefore stand a higher chance of positive reception than a single study. If the replication is not successful, then the PSP should also not matter in the sense that referees' posteriors would also converge, this time towards zero. Whether this scenario warrants a higher chance of publication than a single, statistically significant novel study depends on how much journals value robust null results. Moreover, in case the original study does not replicate, editors should be more open to publishing a null result because of its robustness and because of the added value of such knowledge for the profession. As we gain a firmer understanding of the value of null results, we foresee such robust null results increasing in import (Abadie (2020). A third occurrence is that replications may provide statistically significant results but in opposite directions. Even in this instance, the PSP could guide the interpretation of results, and the results could provide bounds to the effects investigated as well as useful data on how to explore heterogeneity in further studies.

Second, beyond the benefits provided to all authors who care about the robustness of their results, the mechanism described may have a particularly pronounced value for junior scholars. A first reason is that junior scholars typically have limited financial resources to replicate their own work. But the mechanism we investigate here externalizes the cost of replications, providing a comparative advantage to financially constrained scholars.[17] A

[16] It is possible that referees may believe that replications were conducted by sympathetic scholars with a vested interested in successfully replicating, and therefore discount their credibility. As highlighted by Maniadis et al. (2015), this would mean that referees believe that there exists a bias u, generated by "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (Ioannidis (2005), p.697). Referees would update their posteriors as follows: $PSP^{bias} = \frac{(1-\beta)\pi+\beta\pi u}{(1-\beta)\pi+\beta\pi u+[\alpha+(1-\alpha)u](1-\pi)}$. With the presence of the bias u, replications are less effective in moving referees' priors, and therefore referees with very low priors and strong beliefs in the presence of a bias u would update posteriors less than referees believing no bias exists. Still, replicated studies would move posteriors more than a single study, even in presence of the bias u.

[17] A referee noted that the mechanism proposed in this paper formalizes a common occurrence in our profession: Junior scholars offering senior scholars to join their project in exchange for resources and guidance. In this sense, the mechanism reduces the transaction costs associated with finding interested colleagues, and could potentially facilitate such transactions for junior scholars with relatively weaker professional connections.

second reason is that junior scholars may be less likely in successfully placing a novel and surprising study in a highly ranked journal without robust replications. Senior scholars may be more likely to have an established reputation for rigorous scientific conduct, relatively lower pressure from the publish-or-perish culture, perhaps a relatively stronger influence on the editorial process, and larger financial bandwidth to collect additional data when requested by editors.[18] In this manner, this mechanism could help junior scholars establish their reputations by submitting replicated papers to higher ranked journals, with a higher probability of publication. But at the same time, juniors have more time constraints as a result of the tenure system, which may influence their choice of trying to publish their work immediately without seeking replications. It is important to notice that while this mechanism imposes time costs, so does replicating autonomously one's own research. This time cost might be particularly taxing for researchers with limited financial resources, as it takes time to apply for grants. Further, as we highlight in our theory section, short-run impatience may not necessarily lead to faster publications if junior researchers remain too selective in their choice of a target journal.

3 Theoretical Framework: Replications and Time

As discussed in the previous section, a major concern researchers may have in relying on replications is that it is a time consuming process, which may be seen as particularly taxing for researchers with impending time constraints. In this section we present a simple model highlighting under which conditions the replication methodology we investigate in this paper could be chosen over the status quo.

A first order question is whether there are at all incentives for the original authors to engage in costly replications, and how those incentives depend on inter-temporal tradeoffs and editors' preferences. We therefore focus on the decision of the original authors and treat replicators' equilibrium behavior as given. The model however can be easily adapted to study the decision of replicators. More generally, whether there is a nonzero supply to the demand of replicators is an important empirical question that should be investigated

[18] In fact, editors may often be reluctant in asking for additional data to junior scholars, knowing how taxing (and uncertain) this investment could be for them. But the resulting outcome is often that difficult editorial decisions on papers from junior scholars can tilt towards rejection if asking for additional data is perceived by editors as a delicate ask.

in further research. We provide some reflections in section 6.

To characterize the choice of an original author who has to decide whether to invest his time and resources in replications, we employ the job search model developed by DellaVigna and Paserman (2005) (hereinafter DVP), and we adapt it to characterize researchers' publication strategy. The goal of this section is to highlight how the decision to replicate affects the publication process, to highlight which trade-offs scholars with different time horizons face, and the role of editors for promoting replications.

This section delivers two key insights: First, we show that short-run impatience, which is likely experienced by researchers working under tenure-track constraints, may actually increase the length of the publication process: researchers experiencing short-run pressure invest less-than-desired effort in improving their research today, for instance through time-consuming investments such as replications. This reduces the odds of publishing in highly ranked journals. At the same time however, researchers experiencing short-run pressure remain too selective in their publication strategy. The combination of these two forces delays publications. Importantly however, we show that if researchers who experience short-run pressure are sophisticated, they then would be willing to pay for a binding commitment device helping them to invest more in their research and consequently increase their utility. As we show below, the replication mechanism discussed in this paper precisely functions as a binding commitment device.

Second, we show that editors' posture towards replicated work, either successful or unsuccessful, has crucial effects on researchers' willingness to invest in replications.

Model setup

We now describe our adaptation of the DVP model. We refer the reader to the original DellaVigna and Paserman (2005) for proofs.

An infinitely lived researcher who completes a novel working paper in time $t = 0$ faces two sequential decisions: how much effort to invest in improving her paper, and how ambitious to be in terms of academic outlets.

First, the researcher needs to decide how much to invest in the paper: for instance, she can invest today in a number of labor-intensive or capital-intensive technologies that increase the robustness and quality of her research article. Because our paper focuses on one specific type of quality-improving technology – replications – we assume that effort

takes the form of increasing the post-study probability through replications.[19] We thus define this effort level with $r \in [0, 1]$.[20] In each period in which the paper is unpublished, if the researcher exerts effort $r > 0$, she pays a cost of effort $c(r)$ (with $c(r)$ bounded, twice-differentiable, increasing, and strictly convex). We define $c(r)$ as the collection of costs involved in developing the replication strategy, finding researchers willing to replicate, possibly designing ancillary treatments, adapt the original materials to reflect the replication strategy, collect new data, and write up the new paper. Following DVP, we assume that the replication effort r is performed before publication, and it increases the probability of receiving an offer from a desirable academic journal j in any given period t.

Second, a researcher needs to decide how ambitious to be in terms of the publication outlet. Paralleling the definition of reservation wage in the job search literature, we assume that researchers choose a target journal j^*, and skim the publishing opportunities on various journals j accordingly. We define an expected journal offer as either a revise & resubmit invitation, an acceptance, or simply a signal that a given academic outlet is attainable. Desirable journals differ in their quality (or prestige) $j \in [\underline{j}, \bar{j}]$, and a researcher receives an expected offer with probability r.

An expected offer from a desirable journal j is a realization of a random variable J drawn from a known and constant cumulative distribution function $F(j)$, which is independent of replication efforts r.[21] This implies that exerting replication effort r increases the frequency at which an author receives an expected offer from a desirable journal j from $F(j)$, but not the distribution of F.[22] Once an expected offer from journal j is observed, the author must

[19] Note however that any type of costly quality-improving technology fits this framework. For example, any labor-intensive investments (e.g., adding robustness checks, testing alternative models), or capital-intensive investments (e.g., conducting more experiments or treatments, hiring more research assistants, purchasing additional data, replications, etc.).

[20] An effort level close to zero can be thought of as simply submitting a manuscript without replicating.

[21] Since many journals openly report summary statistics about the rates of submissions, acceptances and rejections, this seems like a plausible assumption. Notice that we are implicitly assuming that the likelihood of publishing in prestigious journals is independent of scholars' affiliation or professional networks. We note that skewness in F due to personal connections or affiliations would only strengthen our argument in favor of increasing replication rates.

[22] Note that we implicitly assume that some positive investments in improving the paper are necessary to receive any offer from desirable journals (e.g., if $r = 0$, the researcher receives zero offers). This could fit the case of papers that provide a novel result at odds with the existing literature, or highly exploratory research, which would be difficult to publish on desirable journals absent any investment to improve robustness. As we discussed in previous sections, papers that are well embedded in an established literature, or papers with solid theoretical foundations do not necessarily warrant replications (Maniadis et al., 2017). This can be easily accommodated through a transformation of parameter r. An alternative interpretation is that there exist a set of journals $j < \underline{j}$ that can be accessed without any investments. The value from publishing

make a second decision: whether to accept or reject the offer. Clearly, researchers never refuse acceptance offers from journals they have submitted to. One way to think about expected offers is to assume they are a reduced form of the process of choosing a target desired journal based on signals received from the academic market prior submission (e.g., feedback at conferences, positive signals from editors), where signals about the fit of the paper within journals j's are received with probability r.

Once an offer is accepted in time t, the publication (or R&R) delivers benefits to the author starting from $t+1$. Note that this setup accommodates an occurrence not infrequent in researchers' publication strategy: delaying costly investments, with the hope of obtaining a R&R from a desirable journal.

Time preferences

The main concern that a researcher may have when considering to use replications is that replicating is time consuming, and its benefits accrue only in the future. In particular, authors may differ in how they discount the future. Some researchers, for instance tenured or well-established researchers, may display only a varying degree of long-run impatience – that is – their time preferences may be characterized by exponential discounting. Other researchers, for instance junior or nontenured researchers, may also display short-run impatience, for instance because of the time pressure determined by the tenure-track system. Thus, their time preferences could be described by hyperbolic discounting.

Following DVP, we assume that the present value of a flow of future utilities $(u_t)_{t>0}$ is:

$$u_0 + \beta \sum_{t=1}^{T} \delta^t u_t \tag{2}$$

The discount factor is equal to 1 when $t = 0$, and equal to $\beta\delta^t$ for any subsequent period. When $\beta = 1$, scholars only have long-term impatience (e.g. exponential discounting), and when $\beta < 1$ scholars display both long-term (e.g. δ^t) and short-term (β) impatience (e.g. hyperbolic discounting). Following DVP, we will later distinguish between naive and sophisticated hyperbolic scholars.[23] We now characterize the optimization problem.

in these journals would be easily captured by the value of not publishing in journals $j \in [\underline{j},\bar{j}]$, that is (as shown below) V^U.

[23]Note that in DVP and in traditional time-preferences model, β is usually considered as a preference or trait. In our context, it might be thought as either a context-dependent trait, or an external constraint.

The optimization problem

For all periods t, we define as V_{t+1}^U as the continuation value from having a paper either unpublished or later believed to be unconvincing (e.g., a paper that is rejected in a R&R round, or a paper that is refuted by subsequent research).[24] Similarly, the continuation value at period t from having a paper published in outlet j is $V_{t+1}^P(j)$. The researcher chooses replication effort r and chooses whether to submit to a more or less selective journal j^* to solve the following problem:

$$\max_{r_t \in [0,1]} \; b - c(r_t) + \beta\delta[r_t E_F\{\max(V_{t+1}^P(j), V_{t+1}^U)\} + (1 - rt)V_{t+1}^U] \tag{3}$$

Where b is a time-invariant benefit received at time t. This can be thought as the benefit of circulating and presenting the paper prior publication. At time t the researcher gets benefit b and pays replication cost $c(r_t)$. The continuation payoffs are discounted by a factor $\beta\delta$. With probability r the scholar receives and offer from desirable journal j at time t that he can accept, and obtain from the next period the continuation payoff $V_{t+1}^P(j)$, or reject, thus continuing to receive the continuation value from being unpublished V_{t+1}^U. With probability $1 - r$ the researcher does not receive an offer and continues to receive V_{t+1}^U.

Opportunity cost of replicating

Paralleling the layoff probability proposed by DVP in a labor market context, we allow for the possibility that a published paper (or a paper under revision) is later found to be unconvincing and ultimately rejected or retracted. We characterize this probability as the expected benefits from changing posterior beliefs through replications. We believe that the key value of replication lies precisely in shifting posteriors. We define $B = PSP - \pi = \Delta PSP \in [-1, +1]$ as the expected change in posterior beliefs due to replications, given an initial prior belief $\pi \in [0, 1]$. If replications confirm the robustness of the initial result, then $B > 0$, otherwise $B < 0$ if replications disprove the initial result. We then define $e : [-1, +1] \to [0, 1]$ as a function characterizing the average preferences of journal editors (and possibly referees) for replicated work. The function e maps failed and successful replications in a $[0, 1]$ probability space, and can therefore characterize a wide range of preferences for

[24] We implicitly make the simplifying assumption that a researcher's utility from having a paper unpublished or later found to be unconvincing are the same.

replicated work, either failed and successful.[25] We can then define the probability that a paper is later believed to be convincing as $e(B)$. Consequently the probability that a paper is later believed to be unconvincing is $1 - e(B)$. Thus, the author keeps enjoying a flow of benefits from his paper with probability equal to $e(B)$, the importance editors (and possibly referees) attribute to changes in posterior beliefs due to robust replications.

Following DVP, we consider a stationary environment, and drop the time subscripts. The continuation payoff from a publication or R&R on journal j is therefore:

$$V^P(j) = j + \delta[(1 - e(B))V^U + e(B)V^P(j)] \tag{4}$$

Equation 3 shows that the optimal replication effort and journal acceptance policy depend on the strategies of all future selves through $V^P(j)$ and V^U, and that different selves have conflicting interests, as they would like to delegate the replication effort to the others. Using Markov perfect equilibria as a solution concept, DVP proves that a unique stationary Markov perfect equilibrium exists. The intuition for uniqueness in our setting is the same as in DVP: since replication efforts in the present and in the future are substitutes, and since a feature of Markov perfect equilibria is that strategies do not depend directly on actions that have already been taken, then there are no multiple equilibria in which all researchers' selves either replicate too much or replicate too little. As shown in DVP, researchers accept all journals' expected offers above a certain threshold, and the "reservation journal" in equilibrium is:

$$j^* = (1 - \delta)V^U \tag{5}$$

One can notice that the higher the value for the outside option of remaining unpublished, the higher the researcher aims in terms journal quality. Also, the more long-run impatient a scholar is, the lower is the acceptance threshold. It can be noticed that j^* does not depend on short-run impatience β because benefits from publications or R&R only accrue in future periods, so the threshold for journal quality only depends on long-run impatience δ.

From equations (2) and (3) one can derive the first order conditions with respect to r as a function of the "reservation journal" j^*:

[25] For instance, suppose that $e = |B|$. This would imply that editors values failed and successful replication equally. If instead $e = B$ for $B > 0$ and $e = 0$ otherwise, then editors would only value successful replications.

$$c'(r) = \frac{\beta\delta}{1-\delta(e(B))}[\int_{j^*}^{\bar{j}}(u-j^*)dF(u)] \tag{6}$$

The interpretation of equation (6) straightforwardly follows DVP: at the optimum, the marginal cost of increasing the probability of publishing in a desirable journal through replications equals the marginal benefit, which is the expected present value of publishing in a journal better than j^*.

Equation (5) provides two initial insights.

Observation 1 Increases in short-term impatience (β becomes smaller) decrease the marginal benefit of replicating.

As scholars' short-run impatience increases, the marginal benefit from replicating decreases. This is because the costs of replicating are incurred today, but their benefits only accrue in the future. As shown below, the effect is even stronger for naive hyperbolic discounters, as they expect that their future selves will replicate more than they do, further reducing today's incentives to do so.

Observation 2 Increases in the expected benefits from increasing posteriors through replications $e(B)$ increases the marginal benefit of replications.

Notice that if $e(B)$ is strictly increasing in $B = PSP - \pi$ and $B > 0$, then novel papers that are likely to be met by low priors π have the most to gain from replicating (provided that the result is ultimately true). The value of failed replications depends on the editors' posture towards failed replications that move posteriors to zero. If editors equally value failed and successful replications, for instance if $e(B) = |B|$, then the marginal benefits from replicating are the highest when either priors π are close to zero (and the result is ultimately true), or when priors π are close to one (and the result is ultimately false). That is, researchers have the most to gain from demonstrating through replications that a surprising result is ultimately true, or from demonstrating that a widely-believed result is ultimately false. Observation 2 is the first key takeaway of this section: editors play a fundamental role in determining the value of conducting costly replications. Editors could increase the incentives to replicate more by providing explicit signals about $e(\cdot)$. This is what journals such as Experimental Economics and Journal of the Economic Science Association already do.

Next, we characterize the equilibrium behavior of scholars who differ in their time preferences.

Long-run impatience and the length of the publication process ($\beta = 1, \delta < 1$)

Consider a scholar who only displays long-run impatience (e.g., $\beta = 1, \delta \leq 1$). For instance, one could think of a tenured scholar who does not experience short-run pressure to publish. An increase in long-run patience (i.e., δ increases) has two effects: On the one hand, it increases the marginal benefit from improving the paper through replications. Future benefits become more valuable, and this leads to higher replication efforts, which in turn shortens the time to publish on a desirable journal. On the other hand however, an increase in long-run patience directly increases the researcher's selectivity for accepting a journal of given quality (cf. equation (3)), which increases the time it takes to publish. Thus, more patient scholars both exert more effort to improve their paper and become more selective in their publication strategy. The effect on the publication process is therefore a priori ambiguous. The model of DVP proves (cf. Proposition 4, pp. 538) that the publication rate as a function of δ is hump-shaped if (a) replication efforts become increasingly costly at the margin and (b) the failure rate of publication increases with j (e.g., it is harder to publish on top journals). Both conditions seem quite plausible. Therefore, when researchers become more long-run patient, the replication effort effect dominates the selectivity effect, and the time it takes to publish decreases with patience. On the other hand, when scholars are very patient (e.g., δ closer to one), the selectivity effect dominates the replication effort effect, and the time it takes to publish increases with patience.

Short and long-run impatience and the length of the publication process ($\beta < 1, \delta < 1$)

The presence of both short and long-run impatience generally increases the length of the publication process. As shown earlier, an increase in short-run impatience decreases the marginal benefit from investing in replications which, all else equal, increases the length of the publication process. However, this effect is generally not compensated by a decreased selectivity on j^*.

Consider first hyperbolic researchers who are naive about their short-run impatience: they believe that their future selves will behave as exponential discounters. For naive hyperbolic discounters, an increase in short-run impatience decreases the marginal benefits

of replicating, as the cost of replicating is incurred today but benefits accrue only in the future. This negative effect is reinforced by the fact that naive researchers expect higher effort $c(r)$ from their future selves, further reducing the incentives to invest in replications today. Because naive researchers believe that their continuation payoff from remaining unpublished coincides with the continuation payoff of an exponential discounter (e.g., $V^{U,naive}(\beta,\delta) = V^{U}(\delta)$), then equation (4) implies that naive present biased researchers do not become less selective in their publication strategy j^* as they become more short-term impatient. As a result, an increase in short-run impatience leads to an increase in the length of the publication process if researchers are naive.

Now consider hyperbolic researchers who instead are sophisticated about their present bias. Recall that an increase in short-run impatience for naive researchers has no effect on journal selectivity j^*. For sophisticated researchers, impatience affects j^*, but only indirectly: sophisticated researchers will accept today lower quality publications but only because they know that their future selves will invest too little in replicating. Yet, for sophisticated researchers the overall effect of an increase in short-run impatience on the length of the publication process is still negative as far as the increase in the expected quality of a publication associated with an increase in selectivity j^* (e.g., $\frac{\partial E(J|J\geq x]}{\partial x}$) is smaller than $\frac{1}{1-\beta}$ (Cf. DVP, Proposition 3, pp. 538).[26] For example, if $\beta = \frac{1}{2}$, then $\frac{\partial E(J|J\geq x]}{\partial x}$ should be smaller than 2 for the (sophisticated) selectivity effect to dominate the replication effort effect. In the labor market context, this prediction can be tested using wage data, but in our context we do not have direct means to quantitatively test whether a publication on a top tier journal is, say, twice as better as a publication in a lower ranked journal.[27] Therefore, we do not take a stand on whether the negative effect of short-run impatience on the length of the publication process can be countered by sophisticated researchers.

We instead conclude this section by showing that the possibility of committing to more replications, for instance by using the mechanism described in this paper, can increase the utility of sophisticated researchers who face short-run pressure to publish. As for the naive researchers, an increase in short-run impatience β reduces the marginal benefit of

[26] DVP shows that this condition is always satisfied for a large class of log-concave wage distributions used in the search literature. Further, through a calibration exercise in the labor market context, DVP shows that such sophistication effect on reservation wages is quantitatively small.

[27] One could imagine using the impact factor as a quantitative mean to test this prediction. However, the impact factor of economic journals does not necessarily map directly into their ranking. For instance, the American Economic Review has a lower impact factor compared to other lower ranked journals.

replicating for sophisticated researchers too. However, sophisticated researchers are aware that their future selves will also invest little in replications. Therefore, they would be willing to pay for a commitment device that forces all selves to invest in replications above the equilibrium level $r^{sophisticated}(\beta, \delta)$ determined by equation (4) and (5).

Proposition 1 There exist an $\varepsilon > 0$ such that an increase in the replication effort in all periods from $r^{sophisticated}(\beta, \delta)$ to $r^{sophisticated}(\beta, \delta) + \varepsilon$ strictly increases the net present utility of all the selves of a sophisticated researcher who experiences short-run impatience (cf. DVP, Proposition 1, pp. 537).

Proposition 1 represents the second key takeaway of this section: the presence of short-run impatience hurts the publication prospects of researchers. Junior scholars are most vulnerable to this type of impatience, since they have limited time to secure a tenured position. Sophistication is not enough to guarantee adequate investments in increasing the robustness of a paper, and younger researchers are overall less likely to give up the prospect of better publications: in many academic institutions, a top publication can make or break a young scholar's career. The mechanism proposed in this paper represents a commitment device: a scholar has to commit today to invest in replications. And the mechanism proposed in this paper is designed to make such commitment strictly binding. Proposition 1 shows that such binding commitment can be beneficial to sophisticated researchers who are most exposed to short-run time constraints.

4 Experimental Design and Replication Protocol

We now demonstrate the operation of this mechanism in an experiment on the effects of Knightian uncertainty (or environmental uncertainty) on individual contributions to public goods. This literature is important in its own right, as three key stylized facts have emerged on the private provision of public goods. First, initial contributions to linear public goods typically exceed zero.[28] Second, cooperation decays over time (Andreoni, 1995), a tendency linked to the presence of heterogeneous preferences such as self-interest, altru-

[28] Various factors contribute to higher-than-predicted contributions, such as kindness (Andreoni, 1995), confusion and decision errors (Anderson and Goere, 1998; Houser and Kurzban, 2002), warm-glow (Andreoni, 1990; Palfrey and Prisbrey, 1997), strategic play (Andreoni, 1988), distributional concerns (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), and intentions' signaling (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2004; Cox et al., 2007, 2008).

ism, and (sometimes self-serving) conditional cooperation.[29] Third, centralized institutions such as taxation, competition, and voting rules[30], and decentralized institutions such as communication, moral and monetary sanctioning and rewards[31] contribute to promoting cooperation. In this section, we first introduce our game, then detail the replication procedures and highlight how we contribute to this literature independently of the replication mechanism.

4.1 A Public Goods Game with Environmental Uncertainty

In a standard linear public goods game, participants are randomly assigned to groups of size N. They are endowed with M tokens that they can allocate to a private account that accrues only to their own payoff, or to a group account that pays a Marginal Per Capita Return (MPCR, hereafter) θ to all group members, regardless of their individual contributions. There is no Knightian uncertainty in this game, as θ is perfectly observed by all members. Each player's decision is thus characterized by the following general payoff function:

$$\pi_i = M - g_i + \theta \cdot \sum_{j=1}^{N} g_j \tag{7}$$

with $g_i \in [0,M]$.

We introduce Knightian uncertainty in the public goods game in the following simple way. Instead of observing θ, each participant receives a noisy signal, $s_i = \theta + \varepsilon_i$, where ε_i is distributed according to an unknown distribution, with mean zero and standard deviation σ. It is common knowledge that all signals are drawn from the same distribution. Depending on the treatments, however, participants either observe only their own signal (private signal), or observe their own signal and the signals of all other group members (public signals).

When signals are privately observed, the payoff function takes the form:

[29] See, *e.g.*, Brandts and Schram (2001); Fischbacher et al. (2001); Bowles and Gintis (2002); Frey and Meier (2004); Fischbacher and Gaechter (2010); Ambrus and Pathak (2011); Fischbacher et al. (2014).

[30] See, *e.g.*, Falkinger et al. (2000); Kosfeld et al. (2009); Reuben and Tyran (2010); McEvoy et al. (2011); Putterman et al. (2011); Kesternich et al. (2014).

[31] See, *e.g.*, Fehr and Gaechter (2000); Masclet et al. (2003); Bochet et al. (2006); Sefton et al. (2007); Gaechter et al. (2008); Bochet and Putterman (2009); Nikiforakis (2010).

$$\mathbf{E}[\pi_i] = M - g_i + \mathbf{E}[\theta|s_i] \cdot \sum_{j=1}^{N} g_j \tag{8}$$

When signals are publicly observed instead, the payoff function becomes:

$$\mathbf{E}[\pi_i] = M - g_i + \mathbf{E}[\theta|s_i \bigcap \mathbf{s}_j] \cdot \sum_{j=1}^{N} g_j \tag{9}$$

where $s_i \bigcap \mathbf{s}_j$ is the intersection between a player's own signal and the vector of signals $\mathbf{s}_j$ received by the other group members. This simply means that the true θ has to be compatible with all signals. Equation 9 shows that public signals can vary in how informative they are about the underlying value of θ: If at least two group members receive opposite extreme signals, then θ is perfectly identified and uncertainty is fully resolved. The opposite situation is when $s_j = s \; \forall j$ (*e.g.*, everyone receives the same signal), in which case observing others' signals does not add any useful information.

Let us describe the general procedure, which follows the literature, before providing details about the treatments. In each session, 16 participants play four repeated public goods games in groups of four players. Each game consists of eight rounds. In each round, participants choose how to allocate 10 tokens between a private account and a group account.[32] Each token placed in the private account is worth one token only to the subject. At the end of each round, participants are informed about their own payoff for that round, but are not told how many tokens other players have invested in the group account. After each game, groups are reformed randomly, using a stranger matching procedure. Participants are only identified by a randomly generated ID number. It is common knowledge since the beginning that only one of the four games will be randomly selected for payment, and that each player will be paid the sum of earnings made in the eight rounds that constitute that game.

In all treatments, the instructions specify the possible values of the MPCR. The minimum possible value of the MPCR is 0.05 and the maximum is 1.25, with increments of 0.1. In all treatments, subjects are told that in three out of four games, the MPCR is constant within each game, whereas in one of the four games it is randomly drawn every round (with replacement). In all treatments, the three games with constant MPCR always have the following (predetermined) MPCR values: 0.25, 0.55, and 0.95. There are two sessions per

[32] 20 tokens are worth U.S. $1.

treatment and the order in which games are played is either 0.25, 0.55, 0.95, Variable, or 0.95, 0.55, 0.25, Variable. Variable is always played last, as it is more complex. Before the beginning of each game, participants are informed about whether the game has a constant or variable MPCR.

The experiment consists of four treatments in addition to the baseline treatment. The baseline treatment, *Baseline VCM*, is a standard public goods game without Knightian uncertainty. In two private signal treatments participants only observe their own signals. In the *Private Thin* treatment each participant receives a private signal known to be drawn from the interval: true MPCR ± 0.1. For instance, if a participant receives a private signal of 0.55, they know that the true MPCR can either be 0.45, 0.55, or 0.65. They also know that if the true MPCR is, for instance, 0.65, another player might have received a signal of 0.55, 0.65, or 0.75. In contrast, in the *Private Thick* treatment, participants receive a private signal known to be drawn from the interval: true MPCR ± 0.2. For instance, if a participant receives a private signal of 0.55, they know that the true MPCR can either be 0.35, 0.45, 0.55, 0.65, or 0.75. The two public signals treatments, *Public Thin* and *Public Thick*, have the same parameters as the private conditions, but they differ in the fact that participants also observe the signals of other group members. In the three constant MPCR games, participants receive only one signal per game, whereas in the Variable condition signals are drawn in each new round. Finally, at the end of the experiment in all treatments participants play incentivized tasks to elicit their attitudes toward risk, using the Eckel-Grossman procedure (Eckel and Grossman, 2008), and toward ambiguity.

4.2 Replication Details

The original experiment was conducted at the ExCEN experimental laboratory at Georgia State University, and was programmed using O-Tree (Chen et al., 2016). As specified in the preregistration at the AEA RCT registry and in the original working paper, we conducted a total of three independent replications of the original experiment. A first replication was conducted at the GATE-Lab in Lyon, France. A second replication was conducted at the ICES lab at George Mason University, United States. A third replication was conducted at the MonLEE lab at Monash University, Australia.

The number of replications required was calculated by the original authors based on the results from the original study. The original authors assumed no bias u, and used the significance $\alpha = 0.05$ found in the original study to calculate the PSP. Assuming a prior of

$\pi = 0.01$, they calculated that four total studies (all successfully replicated) would generate a PSP of 0.72 for power equal to 80%.[33] Then, the original authors invited coauthors for the second paper through their professional network.[34]

Each replication closely followed the protocol used in the original experiment, including utilizing the same sample size, the same software, and the same instructions.[35] In total, 640 subjects participated in the experiment (160 in the original study and in each replication, equally balanced across treatments).[36] All subjects were students in local universities. Table A1 in the Appendix shows a balance table for gender and age composition. To account for cross-country differences, the original payoffs were converted into local currencies (France and Australia) and adjusted to reflect the same purchasing power of the original investigation in Atlanta, Georgia. In addition to their payoffs in the game, participants received a show-up fee of $10. On average, they earned US$23.

5 Experimental Results

We organize our results as follows. We first provide summary statistics and nonparametric estimates of the effect of Knightian uncertainty on average contributions in our four experiments across all possible values of the public good (*i.e.*, MPCR). We then analyze our data using an econometric analysis that takes into account group-specific and individual-specific dynamics. Finally, following Maniadis et al. (2015), we calculate the Post Study Probability from the reduced-form estimates of the four experiments, and use the PSP to draw Bayesian inferences about the role uncertainty plays in public goods contributions.

[33] The PSP was calculated to be equal to 0.99 assuming power equal to 50%, as would be the case if only average individual observations were used for results.

[34] After the first three coauthors accepted on a first-come first-serve basis, the original working paper was updated and registered at the AEA RCT registry according to the procedure to reflect these changes. The paper was then circulated online as an NBER working paper. After publication on NBER, other scholars reached out to the original authors to express interest in participating in the project. Given that the project was already registered with the names of the three replication teams, the original authors decided to decline these additional requests to remain true to this initial proof of concept.

[35] One attractive alternative would have been to preregister replications with larger sample sizes (Camerer et al., 2018). We decided to maintain the sample size constant for simplicity, given the exploratory nature of this study, but we do believe larger sample sizes would be highly beneficial.

[36] For the replication conducted in France, the instructions and software materials were translated in French, and translations were independently checked.

5.1 Summary Statistics

Figure 4 and Table 1 provide a first overview of the effect of Knightian uncertainty across the original experiment and the three replications. Each panel of Figure 4 plots the average percentage of the endowment contributed by round for the *Baseline VCM*, the *Private Thin*, and the *Private Thick* treatments across levels of MPCR in the four experiments. Table 1 reports the average percentage of the endowment contributed in each treatment and sample, as well as nonparametric tests (Wilcoxon Mann-Whitney tests, MW hereafter) of the difference between average baseline contributions and contributions in each treatment with Knightian uncertainty, both with private and with public signals.[37]

Together, these results show a mixed effect of uncertainty on cooperation. In the initial study (GSU), the presence of Knightian uncertainty had weak effects on cooperation when the MPCR was equal to 0.25.[38] In contrast, it increased average contributions when the MPCR was equal to 0.55 (average increase relative to *Baseline VCM* of 7.4%, $p < 0.001$, and 4.1%, p=0.07 in *Private Thin* and *Private Thick*, respectively) or equal to 0.95 (average increase of 12%, $p < 0.001$, and 9.1%, $p < 0.01$, in *Private Thin* and *Private Thick*, respectively).

Overall, the initial investigation using GSU data showed a positive effect of Knightian uncertainty on cooperation, which increased with the value of the public good. Figure 4 provides preliminary visual insights about our three replications: First, the GMU sample shows a positive effect of uncertainty on cooperation, which is directionally consistent with our original sample (GSU). Second, the GATE sample has a pattern of cooperation that is inconsistent with our original sample, displaying a mostly null or negative effect of uncertainty. Third, the Monash sample reveals mixed evidence.

The three replications also show heterogeneous effects of uncertainty across different values of the public good. We first look at periods with MPCR equal to 0.25. For the GATE sample, we find a non-significant decrease in average contributions for *Private Thin* of 1.6% (p=0.762), and a significant but small increase of 0.8% for *Private Thick* (p=0.054). By contrast, for the GMU and Monash samples, we find a strong and positive effect of Knightian uncertainty on cooperation. For the GMU sample, average contributions are 8.4% ($p < 0.001$) and 6.5% ($p < 0.001$) higher in *Private Thin* and *Private Thick* relative to

[37] Tables A2, A3, A4 in Appendix provide detailed summary statistics by round and MPCR.

[38] We found a marginally significant increase in average contributions in *Private Thin* relative to *Baseline VCM* equal to 4.1% (p=0.07), and an insignificant average decrease of 2.6% (p=0.3) in *Private Thick* relative to *Baseline VCM*.

Baseline VCM; similarly, for the Monash sample, we find that average contributions are 11.6% ($p < 0.001$) and 8.1% ($p < 0.001$) higher in *Private Thin* and *Private Thick* treatments than in *Baseline VCM*.

For the case of MPCR equal to 0.55, the GMU sample results are consistent with our original study, while the GATE and Monash samples are disparate. In the GMU sample, for example, average contributions in the *Private Thin* treatment are 14.6% higher than *Baseline VCM* ($p < 0.001$) and 5.1% higher (p=0.106) in *Private Thick* than *Baseline VCM*. By contrast, in the Monash sample, Knightian uncertainty has no significant effect in *Private Thin* relative to *Baseline VCM* (an increase of 1.3%, p=0.61), while it significantly reduces contributions in *Private Thick* (a decrease of 9.8%, p=0.005). Similar to Monash, the GATE sample shows a negative effect of Knightian uncertainty on cooperation: average contributions are 12.9% lower in *Private Thin* than in *Baseline VCM* ($p < 0.001$) and 11% lower in *Private Thick* ($p < 0.01$).

Finally, we examine the case of MPCR equal to 0.95. For two out of three replication studies, GMU and Monash, the effect is directionally similar to our original study, but mostly insignificant at conventional levels. Likewise, the GATE sample shows insignificance, but in this case we find a negative effect. For the GMU sample, uncertainty has an insignificant positive effect of 4% (p=0.208) in *Private Thin*, and a significant positive effect of 6.5% (p=0.041) in *Private Thick* relative to the *Baseline VCM*. For the Monash sample, the effect is positive but not statistically significant for both *Private Thin* (3.1%, p=0.615) and *Private Thick* (5.1%, p=0.176) relative to the *Baseline VCM*.

5.2 Econometric Analysis

Thus far, we have abstracted from the fact that in each sample, individuals are repeatedly observed over time t (32 rounds) and make decisions in four separate groups g (eight sequential decisions in each group). To account for these differences, we follow the same econometric strategy used to analyze data in the original study (Butera and List, 2017). For each set of results, we estimate linear models with standard errors clustered both at the group level and at the individual level, as well as linear models with both individual and group fixed effects (see, *e.g.*, Cameron et al. (2008); Correia (2017)).

Empirical results are reported in Table 2 and provide several insights.[39] First, our

[39]For robustness, in online Appendix A we also report coefficient estimates from random effects panel tobit models with group dummies to account for censoring. Left censoring in GSU, GATE, GMU and Monash samples is, respectively: 17.93%, 35%, 20.8%, and 30.43% of the observations.

public treatments provide a useful test for confusion. If participants failed to understand the experimental procedures, then contributions in our public treatment groups in which public signals fully resolve uncertainty should differ from the *Baseline VCM* treatment. For instance, this could happen if subjects failed to take into account other members' signals, or did not understand that the actual MPCR must be compatible with the signals received by all participants.

Table 2 shows that this is not the case. We compare, for each sample, round contributions in the *Baseline VCM* treatment, where the MPCR is known, and round contributions in the two public treatments in which public signals fully resolve uncertainty. Conditional on receiving fully informative public signals ("Fully informative public signals "), contributions are statistically indistinguishable from those in the *Baseline VCM* treatment.[40] This is a useful robustness test to understand how to interpret this set of results.

We next turn to the estimates of the overall effect of Knightian uncertainty. Model 1 in Table 3 reports coefficient estimates from the following model:

$$y_{ig} = \alpha + \beta_1 \mathbf{T} + \beta_2 \mathbf{X}_i + \beta_3 \mathbf{Y}_g + \beta_4 \mathbf{X}_i \cdot \mathbf{T} + \beta_5 \mathbf{X}_g \cdot \mathbf{T} + \varepsilon_{ig} \tag{10}$$

where the dependent variable, y_{ig}, is the contribution to the public goods made by participant i in group g. T is the treatment: *Baseline VCM vs.* Private Signal treatments. $\mathbf{X}_i$ is a vector of individual information, such as the type of signal received. $\mathbf{X}_g$ is a vector of group characteristics, including the value of the MPCR for that group, the contributions made by the other group members, and the types of signals received by the other group members.[41]

The first two columns of Table 3 show the effect of Knightian uncertainty in our original investigation: while initial cooperation levels are not affected, cooperation decays less over time in the presence of uncertainty (variable "Uncertainty X Round number"). In our linear specification with two-way clustered standard errors (model 1), the effect of Knightian

[40] In our original study, we also found that contributions in public treatments marginally increased with the number of MPCR values compatible with the set of public signals. That is, as public signals became less informative, people (marginally) contributed more. We found that contributions increased by 1.081 tokens for each additional admissible value of the MPCR (p=0.068). As detailed in Table A5 in Appendix, the effect of public signals' "informativeness" is not statistically significant for the GMU and Monash samples, whereas it is significant for the GATE sample, although in the opposite direction of our original study: contributions decreased by 1.411 tokens for each additional admissible value of the MPCR (p=0.015).

[41] For model 2 in Table 3, our two-way fixed effects specification, the equation takes the following form: $y_{ig} = \alpha + \beta_1 \mathbf{T} + \beta_2 \mathbf{X}_i + \beta_3 \mathbf{Y}_g + \beta_4 \mathbf{X}_i \cdot \mathbf{T} + \beta_5 \mathbf{X}_g \cdot \mathbf{T} + \eta_i + \gamma_g + \varepsilon_{ig}$, where η_i is an individual fixed effect and γ_g is a group fixed effect.

uncertainty equals 0.078 token per round (p=0.022) while cooperation overall decreases by 0.26 token per round (variable "Round number", $p < 0.001$) – a decrease in the rate of decay of cooperation of about 30%. The effect is larger, albeit only marginally significant, under our two-way fixed effects specification (42%, p=0.081), and equal to 39% in our panel tobit specification (p=0.024, see Appendix A). At odds with the original data, the decay of cooperation is not statistically different in the presence of uncertainty in any of the replication samples at conventional levels.[42] There is a statistically significant effect of uncertainty in the Monash sample only in model 1 ($p < 0.01$), mostly driven by high initial contributions in the Private Signal treatments relative to *Baseline VCM* for the period with MPCR equal to 0.25.

5.3 Bayesian Analysis of Replications

The headline result in the original BL study was that Knightian uncertainty increased cooperation in public goods games, suggesting interesting implications for private provision of public goods in the field. This struck us as a foundational result. We can now ask, with these new data, does the presence of Knightian uncertainty effectively increase cooperation in public goods games? Our non-parametric and econometric results provide mixed evidence, hinting at a positive effect of uncertainty in reduced-form estimates for the GMU sample and in econometric estimates for the Monash sample, and hinting at a null effect for the GATE sample (and a negative effect in reduced-form estimates for the MPCR equal to 0.55).

In this section, we assess how a Bayesian would update their beliefs about the overall effect of Knightian uncertainty on cooperation after observing the initial results and three replications. We do so for different possible initial priors to showcase how a few replications, both successful and failed alike, can allow robust convergence of Post-Study Probabilities and facilitate the advancement of economic science. We focus on reduced-form estimates of the overall effect of uncertainty (*Baseline VCM vs. Private Signal* treatments). We conservatively compare average individual contributions. For each sample, this results in 32 observations in the *Baseline VCM*, and 64 observations in the *Private Signal* treatments.

To conduct our Bayesian analysis we follow the approach of Maniadis et al. (2015). Let each researcher's study have the same power $(1 - \beta)$. The probability that at least one of the k researchers will declare a true association as true is $(1 - \beta^k)$. Likewise, the

[42] The same holds for our panel tobit specification, see Table A6 in Appendix.

probability that a false relationship is declared true by at least one of k researchers is $1-(1-\alpha)^k$. Hence, in the presence of competition by independent researchers the Post-Study Probability PSP^{comp} is equal to:

$$\text{PSP}^{comp} = \frac{(1-\beta^k)\cdot\pi}{(1-\beta^k)\cdot\pi + [1-(1-\alpha^k)](1-\pi)} \tag{11}$$

Table 4 reports the average of the individual contributions in the four samples. In the original BL sample, average individual contributions were overall 7% higher (0.7 tokens) in our *Private Signal* treatments than in the *Baseline VCM* treatment (p=0.054). This corresponds to a 0.41 standard deviation increase in contributions due to Knightian uncertainty. The *ex-post* power $(1-\beta)$ for such reduced-form result is therefore equal to 50%. Using this conservative test, none of the three replication samples show a statistically significant effect of uncertainty on cooperation. We can therefore use equation 11 to compute the PSP. Table 5 provides an overview of the PSP given different possible priors π, after our initial (significant) study and after our three (failed) replications.

Table 5 conveys three critical messages. First, small deviations in priors π cause large differences in posteriors after a single, successful investigation. For instance, Column 1 in Table 5 shows that with priors $\pi = 0.01$, the PSP increases to 0.09 after our initial successful study. However, with slightly higher priors, for instance $\pi = 0.1$, the PSP would notably increase to 0.53 after this first study. Second, after a single successful study, it is very likely for the PSP to be higher than 0.5 for a wide range of priors. In our case, as highlighted in bold in column 1, the PSP is strictly greater than 0.5 for priors $\pi \geq 0.1$. Third, and most importantly, a few replications allow posteriors to converge. Column 4 shows that with three replications, posterior beliefs above 0.5 are only generated by large priors $\pi \geq 0.5$, which are very unlikely in the context of novel and surprising findings, as is the case in our study.

6 Discussion and Conclusion

This paper analyzes a novel, voluntary mechanism to promote replications within the sciences. The mechanism is not intended to replace existing approaches, but rather to enrich the set of tools available in the profession to promote the replication of research results. We demonstrate the functioning of this mechanism with an investigation of the effects of Knightian uncertainty (ambiguity) on providing money for a privately-provided

public good, a pervasive and yet insufficiently explored feature of such institutions. The original, voluntarily unpublished study (Butera and List, 2017) unexpectedly found that ambiguity about the value of a public good facilitates cooperation. We report results from the original study and three independent replications, and show that while ambiguity has a positive effect in two replications for low-quality public goods, overall the original results do not pass a conservative replication test. We conclude that Knightian uncertainty likely has a limited impact on cooperation, corroborating the existing approach of focusing on strategic uncertainty to study public goods.

This decentralized and "price-driven" mechanism addresses the incentive problem that both original authors and "replicators" typically face. The original authors of a study prefer to publish their novel results without the added cost of replicating, preferably in a highly ranked journal. As Maniadis et al. (2015) point out however, given the mechanics of statistical inference, posterior beliefs based on a single, novel exploration are quite sensitive to initial priors. Because novel and surprising results are likely to face low priors, the successful publication of these studies relies heavily on small variations in the distribution of prior beliefs. A few successful replications, on the other hand, can increase the robustness of novel results by allowing posterior beliefs to converge (Coffman and Niederle, 2015). Even unsuccessful replications, as is the case for the study in this paper, allow beliefs to converge and provide a constructive use for null results. We therefore believe that the approach analyzed in this study may be particularly well-suited for novel studies likely to suffer from low priors, and particularly when conducted by scholars at the early stages of their careers. A positive externality for journal editors is the greater incentive for authors to replicate their findings before initial submission. Furthermore, there is another positive externality for the profession itself, as replications contribute to increase the normative value of economics, and its visibility. Indeed, public and private policy makers would certainly be more likely to take into account recommendations stemming from replicated research.

Clearly, this mechanism is only one of many steps towards promoting a more widespread use of replications in economics, and does not directly address a number of important empirical questions.

First, our current model is silent relative to how to optimally choose the number of replications. The approach for this study was to estimate the *ex ante* PSP – and consequently the number of replications needed – under two assumptions: i) we assumed no bias u in the PSP (neither sympathetic nor antagonistic); ii) we computed the power $(1-\beta)$

based on the results of the main specification in the original study. These assumptions are ex-post innocuous for the current paper, since we conservatively concluded that the original study did not replicate. However, this is not generally true. To see this, suppose that we did successfully replicate. An editor or a referee might have raised doubts about the independence of replications – perhaps due to the fact the original authors and replicators knew each other, or for other reasons. Such concerns do not invalidate the replications per se, but do affect by how much observers update their priors. This would imply that a Bayesian would penalize the PSP by a factor $u > 0$: the ex-post PSP would have then been lower than the PSP calculated ex-ante. Consequently, three replications might have been insufficient to let posteriors converge. Alternatively, an editor or referee might have requested a more conservative approach to data analysis, for instance (as we did in our second paper, this paper) to only compare average individual observations. In this case, the ex-post PSP would have differed from the ex-ante PSP due to reduced power $(1 - \beta)$ of the test used in the second paper. Further research should address this important empirical issue.

Second, we opted for a simple and binary definition of "what it means to replicate". We evaluate our replication studies by whether they generate or not a statistically significant result in the same direction as the original study. In doing so, our approach admittedly sets aside useful information contained in our collection of studies, such as effect sizes and standard errors. To paraphrase an insightful comment from an anonymous referee, we (social scientists) should care about estimating the effect size of a given phenomenon and quantify our uncertainty about it, rather than accumulating knowledge about whether an effect is different from zero. We take a simplified Bayesian approach to put emphasis on the replication mechanism itself, but we note that our approach shares the same limitations of current inference methods (e.g. using p-values as cutoffs). We acknowledge the benefits of richer Bayesian analyses and hope that future studies will embrace such route.[43]

Third, the mechanism described in this paper could be well-suited for relatively young scholars, as it provides them with a better chance to score a stronger publication and establish their reputation. Yet, it remains empirically unclear what supply will look like on the replicators' side. Established scholars may have an interest in betting on young researchers' ideas by providing resources and coauthoring with them. Similarly, their Ph.D. students may join the replication teams to improve their research skills, concretely im-

[43] We are grateful to an anonymous referee for raising these important points.

plement replications, and begin publishing. Alternatively, senior scholars may have their own projects that they would rather fund. Other young researchers working in the same area of research may also be interested in teaming up with their peers. This would allow them to share the costs of research, and share a better chance at stronger publications. Yet, because the paternity of the original idea would be common knowledge, they may be dissuaded and might prefer to focus their effort on other independent ideas. The relative weight that tenure committees place on stronger publications versus stronger reputation for original ideas may differ across institutions, and so might the subjective beliefs young scholars have about those weights. These factors would therefore affect the opportunity cost of joining a replication paper.

Fourth, a widespread adoption of replication mechanisms like the one described here, coupled with increasing replication requirements from editors, could raise concerns about inequality among researchers: at scale, a fear might be that only relatively successful and established scholars would be able to leverage enough interest in their work to replicate and publish in high ranked journals, while other scholars would be left with the role of replicators. Innovations attempting to improve scientific standards may increase barriers to entry. But barriers to entry, especially for young experimental and behavioral economists, already exist and are substantial: laboratory experiments' costs can increase quickly with large sample sizes and increasing subjects' payoffs. Field experiments not only require financial resources, but also organizational resources and connections with companies and institutions that scholars early in their career might not have. As a result, young scholars lacking connections, institutional reputations, and financial resources are *already* approaching the publication market with a handicap. It is crucial that such hurdles are recognized and that steps be taken to attenuate them, and we believe this should be done while simultaneously exploring new procedures to promote a productive and wider use of replications.

Finally, the mechanism analyzed here may pose some implementation challenges in the presence of high fixed costs or organizational and institutional constraints, such as for large-scale field experiments. In some instances, an exact replication of a large scale RCT may simply not be feasible. Two observations can be made in this regard. First, while exact replications may be difficult or impossible, replicating within a different setting or with different parameters could be feasible. As discussed in section 2,replications could for instance be carried out following the Selection, Attrition, Naturalness, and Scaling (SANS) framework proposed by List (2020). Second, and more substantially, a replication mechanism such as the one analyzed here could help promote the implementation of field

experiments that would otherwise be obscure. In some instances, scholars may hesitate to invest time and resources in otherwise viable research projects, perhaps due to the fact that the scale of the experiment is not large enough to provide conclusive answers, or that the available field setting is not entirely policy-relevant relative to the research question at hand. Yet, such initial experiments may be crucial data points, and when combined with further replications they could critically expand the scope and frequency of experimental research.

References

Abadie, A. (2020): "Statistical Nonsignificance in Empirical Economics," American Economic Review: Insights, 2, 193–208.

Abdulkadiroglu, A., J. Angrist, and P. Pathak (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," Econometrica, 82, 137–196.

Abeler, J., D. Nosenzo, and C. Raymond (2019): "Preferences for truth-telling," Econometrica, 87, 1115–1153.

Ambrus, A. and P. Pathak (2011): "Cooperation over finite horizons: A theory and experiments," Journal of Public Economics, 95, 500–512.

Anderson, S. and C. A. Goere, Jacob K.and Holt (1998): "A Theoretical Analysis of Altruism and Decision Error in Public Goods Games," Journal of Public Economics, 70, 297–323.

Andreoni, J. (1988): "Why Free Ride? Strategies and Learning in Public Goods Experiments," Journal of Public Economics, 37, 291–304.

——— (1990): "Impure altruism and donations to public goods: a theory of warm-glow giving," The Economic Journal, 100, 464–477.

——— (1995): "Cooperation in public goods experiments: Kindness or confusion?" American Economic Review, 85, 891–904.

Andrews, I. and M. Kasy (2017): "Identification of and correction for publication bias," Working paper.

——— (2019): "Identification of and correction for publication bias," American Economic Review, 109, 2766–94.

Benjamin, D., J. Berger, M. Johannesson, B. Nosek, E.-J. Wagenmakers, R. Berk, K. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. Chambers, M. Clyde, T. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. Field, M. Forster, E. George, R. Gonzalez, S. Goodman, E. Green, D. Green, A. Greenwald,

J. Hadfield, L. Hedges, L. Held, T.-H. Ho, H. Hoijtink, J. Jones, D. Hruschka, K. Imai, G. Imbens, J. Ioannidis, M. Jeon, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. Maxwell, M. McCarthy, D. Moore, S. Morgan, M. Munaf, S. Nakagawa, B. Nyhan, T. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. Schnbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. Watts, C. Winship, R. Wolpert, Y. Xie, C. Young, J. Zinman, and V. Johnson (2018): "Redefine Statistical Significance," Nature Human Behaviour, 2, 6–10.

Berry, J., L. C. Coffman, D. Hanley, R. Gihleb, and A. J. Wilson (2017): "Assessing the rate of replication in economics," American Economic Review, 107, 27–31.

Bjork, L., M. Kocher, P. Martinsson, and P. Nam Khanh (2016): "Cooperation under risk and ambiguity," Working Paper in Economics, University of Gothenburg, 683.

Bochet, O., T. Page, and L. Putterman (2006): "Communication and Punishment in Voluntary Contribution Experiments," Journal of Economic Behavior & Organization, 60, 11–26.

Bochet, O. and L. Putterman (2009): "Not just babble: Opening the black box of communication in a voluntary contribution experiment," European Economic Review, 53, 309–326.

Bolton, G. E. and A. Ockenfels (2000): "ERC: A theory of equity, reciprocity, and competition," American Economic Review, 90, 166–193.

Bowles, S. and H. Gintis (2002): "Social Capital and Community Governance," The Economic Journal, 112, F419–F436.

Brandts, J. and A. Schram (2001): "Cooperation and Noise in Public Goods Experiments: Applying the Contribution Function Approach," Journal of Public Economics, 79, 399–427.

Butera, L. and J. A. List (2017): "An Economic Approach to Alleviate the Crisis of Confidence in Science: With an Application to the Public Goods Game," NBER Working Papers, 23335.

Camerer, C., A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, W. E.-J., and H. Wu (2018): "Evaluating the replicability of social science experiments in Nature and Science," Nature Human Behaviour, 2, 637–644.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, et al. (2016): "Evaluating replicability of laboratory experiments in economics," Science, 351, 1433–1436.

Cameron, A., J. Gelbach, and D. Miller (2008): "Bootstrap-based improvements for inference with clustered errors," Review of Economics and Statistics, 90, 414–427.

Cesarini, D., E. Lindqvist, R. Ostling, and B. Wallace (2016): "Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players," The Quarterly Journal of Economics, 131, 687–738.

Charness, G. and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," The Quarterly Journal of Economics, 117, 817–869.

Chaudhuri, A. (2011): "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," Experimental Economics, 14, 47–83.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTreeAn open-source platform for laboratory, online, and field experiments," Journal of Behavioral and Experimental Finance, 9, 88–97.

Coffman, L. C. and M. Niederle (2015): "Pre-analysis plans have limited upside, especially where replications are feasible," Journal of Economic Perspectives, 29, 81–97.

Coffman, L. C., M. Niederle, and A. J. Wilson (2017): "A Proposal to Organize and Promote Replications," American Economic Review: Papers and Proceedings, 107, 41–45.

Correia, S. (2017): "Linear models with high-dimensional fixed effects: An efficient and feasible estimator," Mimeo.

Cox, J. C., D. Friedman, and S. Gjerstad (2007): "A tractable model of reciprocity and fairness," Games and Economic Behavior, 59, 17–45.

Cox, J. C., D. Friedman, and V. Sadiraj (2008): "Revealed Altruism," Econometrica, 76, 31–69.

de Ree, J., K. Muralidharan, M. Pradhan, and H. Rogers (2018): "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia," The Quarterly Journal of Economics, 133, 993–1039.

DellaVigna, S. and E. Linos (2022): "RCTs to scale: Comprehensive evidence from two nudge units," Econometrica, 90, 81–116.

DellaVigna, S. and M. D. Paserman (2005): "Job search and impatience," Journal of Labor Economics, 23, 527–588.

Drazen, A., A. D. Almenberg, E. Y. Ozbay, and E. Snowberg (2019): "A Journal-Based Replication of Being Chosen to Lead," NBER Working Papers, No. w26444.

Drazen, A. and E. Y. Ozbay (2019): "Does being chosen to lead induce non-selfish behavior? Experimental evidence on reciprocity," Journal of Public Economics, 174, 13–21.

Dreber, A., d. A. J. Pfeiffer, Thomas a, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson (2015): "Using Prediction Markets to Estimate the Reproducibility of Scientific Research," Proceedings of the National Academy of Sciences, 112, 15343–15347.

Dufwenberg, M. and G. Kirchsteiger (2004): "A theory of sequential reciprocity," Games and Economic Behavior, 47, 268–298.

Eckel, C. C. and P. J. Grossman (2008): "Forecasting risk attitudes: An experimental study using actual and forecast gamble choices," Journal of Economic Behavior & Organization, 68, 1–17.

Falkinger, J., E. Fehr, S. Gaechter, and R. Winter-Ebmer (2000): "A Simple Mechanism for the Efficient Provision of Public Goods: Experimental Evidence," American Economic Review, 90, 247–264.

Fehr, E. and S. Gaechter (2000): "Cooperation and Punishment in Public Goods Experiments," American Economic Review, 90, 980–994.

Fehr, E. and K. Schmidt (1999): "A Theory of Faimess, Competition, and Cooperation," The Quarterly Journal of Economics, 114, 817–868.

Fischbacher, U. and F. Föllmi-Heusi (2013): "Lies in disguisean experimental study on cheating," Journal of the European Economic Association, 11, 525–547.

Fischbacher, U. and S. Gaechter (2010): "Social preferences, beliefs, and the dynamics of free-riding in public good experiments," American Economic Review, 100, 541–556.

Fischbacher, U., S. Gaechter, and E. Fehr (2001): "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment," Economics Letters, 71, 397–404.

Fischbacher, U., S. Schudy, and S. Teyssier (2014): "Heterogeneous reactions to heterogeneity in returns from public goods," Social Choice and Welfare, 43, 195–217.

Fischer, R. (1935): The Design of Experiments, Edinburgh: Oliver and Boyd.

Fisher, J., R. Isaac, J. Schatzberg, and J. Walker (1995): "Heterogenous Demand for Public Goods: Behavior in the Voluntary Contributions Mechanism," Public Choice, 85, 249–266.

Franco, A., N. Malhotra, and G. Simonovits (2014): "Social science. Publication bias in the social sciences: unlocking the file drawer," Science, 345, 1502–1505.

Fréchette, G. R., K. Sarnoff, and L. Yariv (2021): "Experimental Economics: Past and Future," .

Frey, B. S. and S. Meier (2004): "Pro-social behavior in a natural setting," Journal of Economic Behavior & Organization, 54, 65–88.

Gaechter, S., E. Renner, and M. Sefton (2008): "The long run benefits of punishment," Science, 322, 1510.

Gangadharan, L. and V. Nemes (2009): "Experimental analysis of risk and uncertainty in provisioning private and public goods," Economic Inquiry, 47, 146–164.

Hamermesh, D. S. (2007): "Viewpoint: Replication in economics," Canadian Journal of Economics, 40, 715–733.

——— (2017): "Replication in Labor Economics: Evidence from Data, and What It Suggests," American Economic Review, 107, 37–40.

Harrison, G. W. and J. A. List (2004): "Field Experiments," Journal of Economic Literature, 42, 1009–1055.

——— (2008): "Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse," The Economic Journal, 118, 822–843.

Houser, D. and D. Kurzban (2002): "Revisiting Kindness and Confusion in Public Goods Experiments," American Economic Review, 92, 1062–1069.

Ioannidis, J. (2005): "Contradicted and initially stronger effects in highly cited clinical research," Journal of the American Medical Association, 294, 218–228.

Kessler, J. B. and S. Meier (2014): "Learning from (failed) replications: Cognitive load manipulations and charitable giving," Journal of Economic Behavior & Organization, 102, 10–13.

Kesternich, M., A. Lange, and B. Sturm (2014): "The Impact of Burden Sharing Rules on the Voluntary Provision of Public Goods," Journal of Economic Behavior & Organisation, 105, 107–123.

Kosfeld, M., A. Okada, and A. Riedl (2009): "Institution Formation in Public Goods Games," American Economic Review, 99, 1335–1355.

Landy, J., M. Jia, I. Ding, D. Viganola, W. Tierney, A. Dreber, M. Johanneson, T. Pfeiffer, C. Ebersole, Q. Gronau, et al. (2020): "Crowdsourcing hypothesis tests: Making transparent how design choices shape research results," Psychological Bulletin, 146, 451–479.

Ledyard, J. (1995): "Public goods: A survey of experimental research," in The Handbook of Experimental Economics, ed. by J. H. Kagel and A. E. Roth, Princeton: Princeton University Press.

Levati, M. V., A. Morone, and A. Fiore (2009): "Voluntary contributions with imperfect information: An experimental study," Public Choice, 138, 199–216.

List, J. A. (2020): "Non est disputandum de generalizability? A glimpse into the external validity trial," Tech. rep., National Bureau of Economic Research.

List, J. A., A. M. Shaikh, and Y. Xu (2019): "Multiple hypothesis testing in experimental economics," Experimental Economics, 22, 773–793.

Maniadis, Z., F. Tufano, and J. A. List (2014): "One swallow doesn't make a summer: New evidence on anchoring effects," American Economic Review, 104, 277–90.

——— (2015): How to make experimental economics research more reproducible: lessons from other disciplines and a new proposal, vol. 18, Cheltenham: Edward Elgar Publishing.

——— (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, F209–F235.

Masclet, D., C. Noussair, S. Tucker, and M. C. Villeval (2003): "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism," American Economic Review, 93, 366–380.

McEvoy, D. M., J. J. Murphy, J. M. Spraggon, and J. K. Stranlund (2011): "The problem of maintaining compliance within stable coalitions: experimental evidence," Oxford Economic Papers, 63, 475–498.

Meghir, C., M. Palme, and E. Simeonova (2018): "Education and Mortality: Evidence from a Social Experiment," American Economic Journal: Applied Economics, 10, 234–256.

Moonesinghe, R., M. J. Khoury, and A. C. J. Janssens (2007): "Most Published Research Findings Are False – But a Little Replication Goes a Long Way," PLOS Medicine, 4, e28.

Nikiforakis, N. (2010): "Feedback, punishment and cooperation in public good experiments," Games and Economic Behavior, 68, 689–702.

Nikiforakis, N. and R. Slonim (2019): "Editors Preface: Trends in experimental economics (1975–2018)," Journal of the Economic Science Association, 5, 143–148.

Palfrey, T. and J. Prisbrey (1997): "Anomalous Behavior in Public Goods Experiments: How Much and Why?" American Economic Review, 87, 829–846.

Putterman, L., J.-R. Tyran, and K. Kamei (2011): "Public goods and voting on formal sanction schemes," Journal of Public Economics, 95, 1213–1222.

Rabin, M. (1993): "Incorporating Fairness into Game Theory and Economics," American Economic Review, 83, 1281–1302.

Reuben, E. and J.-R. Tyran (2010): "Everyone is a winner: promoting cooperation through all-can-win intergroup competition," European Journal of Political Economy, 26, 25–35.

Sefton, M., R. Shupp, and J. Walker (2007): "The Effect of Rewards and Sanctions in the Provision of Public Goods," Economic Inquiry, 45, 671–690.

Smith, V. (2018): Learning from Experiments that Fail to Confirm Beliefs: Three Cases, Doctorate Honoris Causa lecture – GATE, Lyon.

Smith, V. L., G. Suchanek, and A. Williams (1988): "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets," Econometrica, 56, 1119–1151.

Theroude, V. and A. Zylbersztejn (2020): "Cooperation in a risky world," Journal of Public Economic Theory, 22, 388–407.

Villeval, M. C. (2020): "Public goods, norms and cooperation," in Handbook of Experimental Game Theory, ed. by M. Capra, R. Croson, M. Rigdon, and T. Rosenblat, Cheltenham: Edward Elgar Publishing, vol. Chapter 7, 184–212.

Wacholder, S., S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman (2004): "Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies," Journal of the National Cancer Institute, 96, 434–442.

Whaples, R. (2006): "The Costs of Critical Commentary in Economics Journals," Journal Watch, 3, 275–282.

Ziliak, S. and D. McCloskey (2008): The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives. Economics, Cognition, And Society, University of Michigan Press.

Figures

Figure 1: Post-Study Probability (PSP) of a Given Result Being True as a Function of the Number of Failed Replications and Priors $\pi = \{1\%, 10\%\}$ *(assuming* $\alpha = 0.05, (1-\beta) = 80\%$*)*

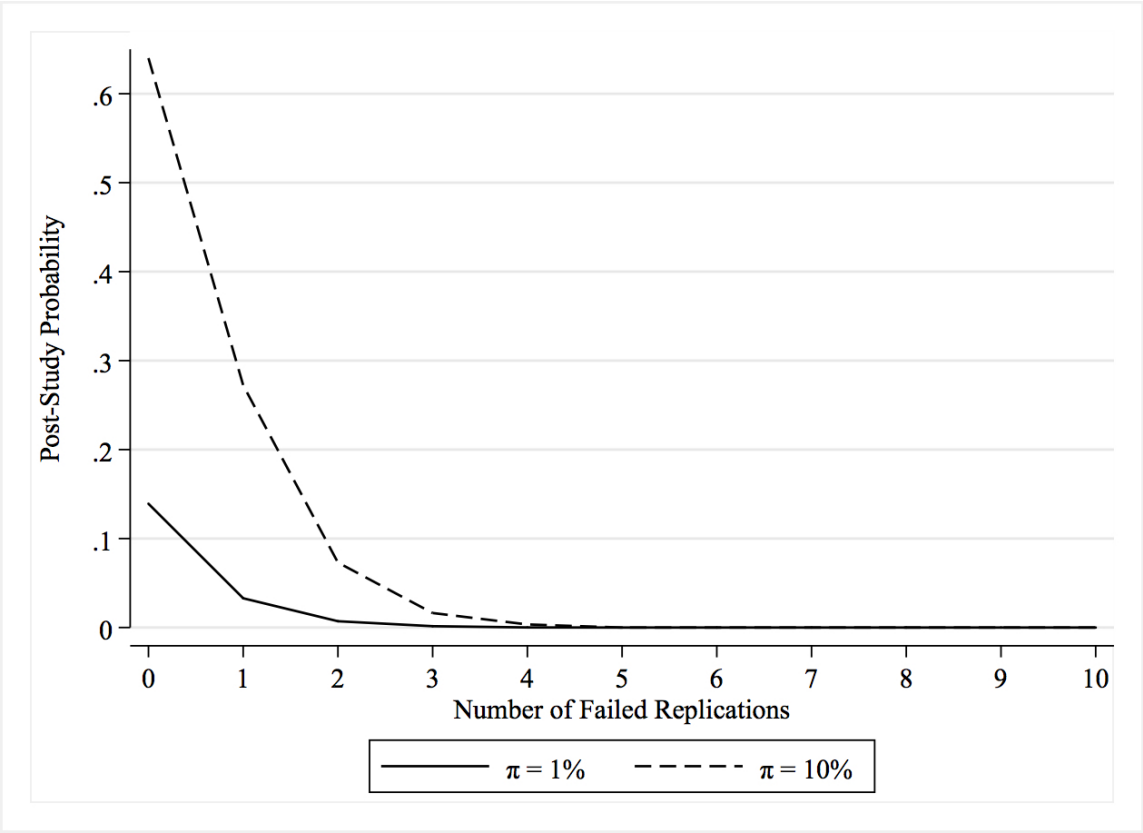


Figure 2: Post-Study Probability (PSP) of a Given Result Being True as a Function of the Number of Successful Replications (assuming $\pi = 0.01, \alpha = 0.05, (1 - \beta) = 80\%$)

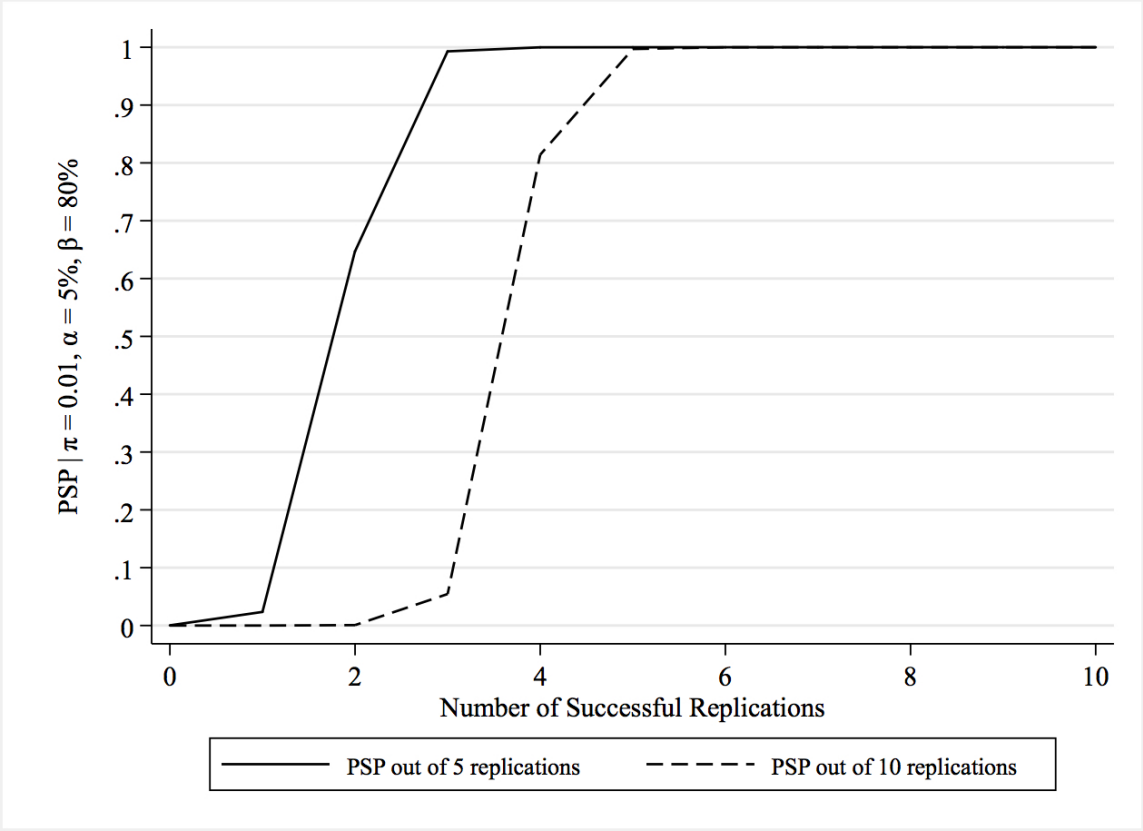


Figure 3: Post-study probability as a function of n. of successful replications (assuming $\pi = 0.01$)

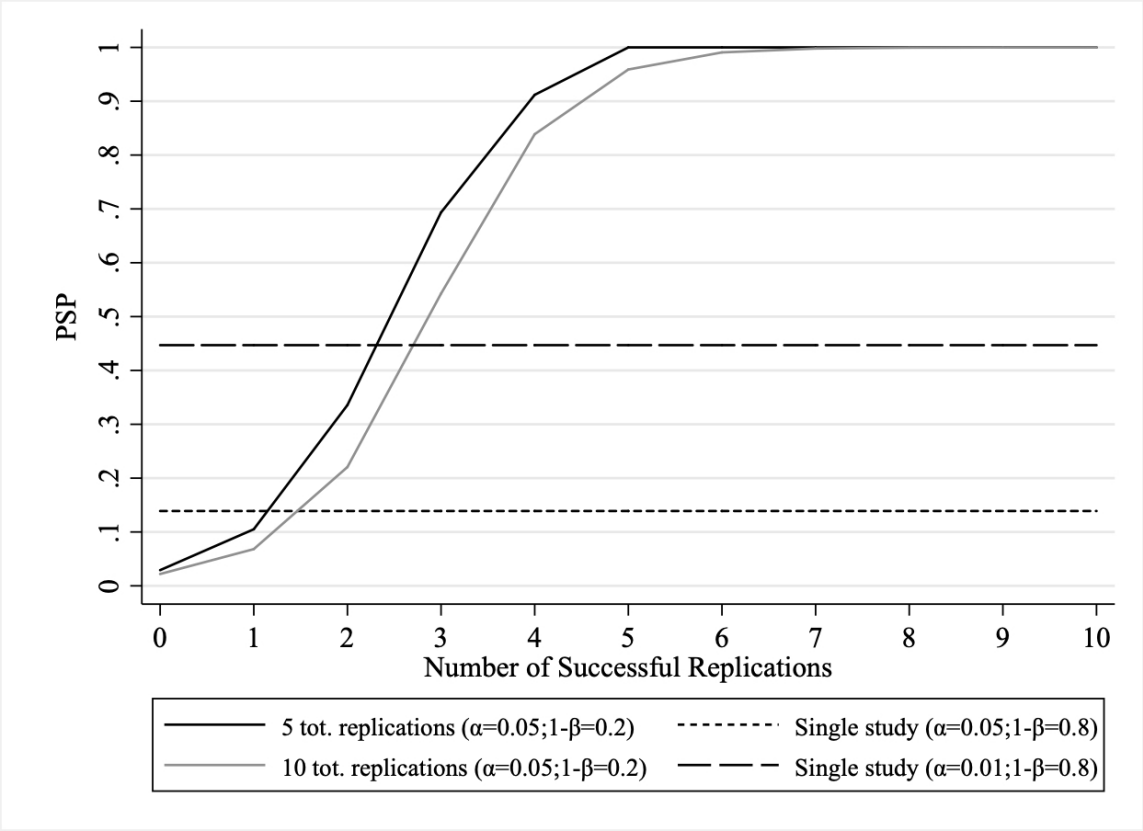
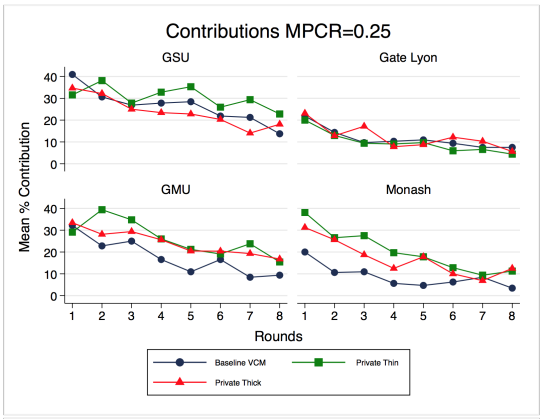
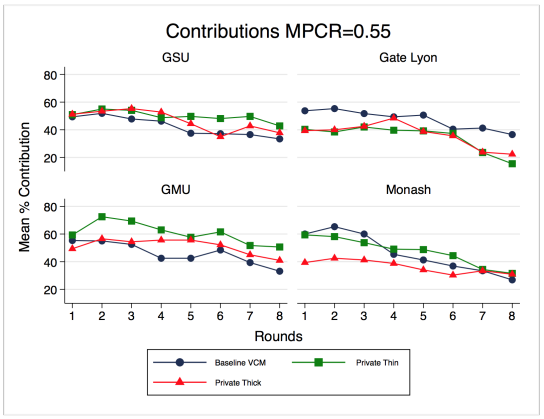
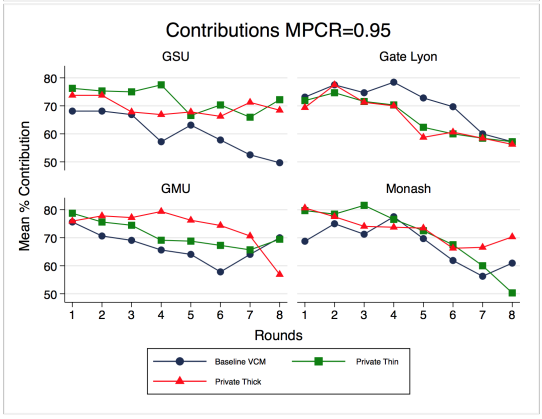


Figure 4: Average contributions (%) by round, MPCR, sample and treatments







Tables

Table 1: Average Contributions as Percentage of Endowment, by Treatment and Location

	GSU			GMU		
MPCR	0.25	0.55	0.95	0.25	0.55	0.95
	Avg. %	Avg. %	Avg. %	Avg. %	Avg. %	Avg. %
Baseline VCM	26.4	42.5	60.4	17.7	46.1	67.1
Private Thin	30.5	49.9	72.4	26.1	60.7	71.1
Private Thick	23.8	46.6	69.5	24.2	51.2	73.6
Public Thin	25.4	41.6	65.3	23.4	55	74.5
Public Thick	30.3	43.8	65.2	29.1	55.9	72.7
Baseline - Private Thin	-4.1*	-7.4***[†]	-12***[††]	-8.4***[†]	-14.6***[††]	-4 ns
Baseline - Private Thick	2.6 ns	-4.1*	-9.1***[††]	-6.5***	-5.1 ns	-6.5 **
Baseline - Public Thin	1 ns	0.9 ns	-4.9 ns	-5.7*	-8.9***[††]	-7.4 ns [†]
Baseline - Public Thick	-3.9 **	-1.3 ns	-4.8 ns	-11.4***[††]	-9.8***[†]	-5.6 ns [†]

	Monash			GATE		
MPCR	0.25	0.55	0.95	0.25	0.55	0.95
	Avg. %	Avg. %	Avg. %	Avg. %	Avg. %	Avg. %
Baseline VCM	8.8	46.1	67.7	11.4	47.4	70.4
Private Thin	20.4	47.4	70.8	9.8	34.5	65.8
Private Thick	16.9	36.3	72.8	12.2	36.4	65.3
Public Thin	12.4	41.6	67.9	9.9	41.1	64.2
Public Thick	15.8	48.1	72	11.8	34.3	61.9
Baseline - Private Thin	-11.6***[††]	-1.3 ns	-3.1 ns	1.6 ns	12.9***[††]	4.6 ns
Baseline - Private Thick	-8.1***[†]	9.8***	-5.1 ns	-0.8*	11***[††]	5.1 ns
Baseline - Public Thin	-3.6***	4.5 ns	-0.2 ns	1.5 ns	6.3*	6.2 ns
Baseline - Public Thick	-7***[†]	-2 ns	-4.3*	-0.4 ns	13.1***[††]	8.5*[†]

Notes: Table 1 reports average contributions expressed as a percentage of the endowment for our four different samples. Contributions are averaged by treatment and by MPCR (Marginal Per Capita Return – or quality of the public goods). For each sample, the last four rows report the percentage difference in contributions between the baseline and each treatment. The pairwise treatment comparisons are based on two-tailed Mann-Whitney tests. ns: not significant, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. We report in brackets the p-values corrected for multiple hypothesis testing using the procedure from List et al. (2019). ns: no dagger, † $p < 0.10$, †† $p < 0.05$, ††† $p < 0.01$.

Table 2: Determinants of the Effect of Fully Informative Public Signals on Contributions to the Public Goods, by Location

	GSU	GATE	GMU	Monash
Model	(1)	(2)	(3)	(4)
Dependent variable	Contribution	Contribution	Contribution	Contribution
Fully informative public signals	0.326	-0.632	0.71	0.105
(0=Baseline VCM; 1=yes)	(0.369)	(0.472)	(0.477)	(0.489)
Round number (1 to 8)	-0.229***	-0.208***	-0.234***	-0.260***
	(0.036)	(0.034)	(0.038)	(0.031)
Period (1 to 4)	-0.262***	-0.0282	-0.208	-0.156
	(0.097)	(0.084)	(0.138)	(0.126)
Order (1= 0.25, 0.55, 0.95, Var.;	0.702*	-0.288	-0.864*	-0.088
2= 0.95, 0.55, 0.25, Var.)	(0.373)	(0.488)	(0.480)	(0.495)
Value of MPCR	5.745***	8.242***	6.577***	7.897***
	(0.489)	(0.467)	(0.516)	(0.535)
Number of observations	1,960	1,992	1,888	2,000
R-squared	0.293	0.402	0.304	0.351

Notes: The models report estimates from linear models with standard errors clustered both at the group and individual levels. The data only includes observations from the *Baseline VCM* treatment and from groups within the *Public Signals* treatments (both *Thin* and *Thick*) in which public signals uniquely identify the true MPCR. "Value of MPCR" identifies the true MPCR for the round. Note that in any given period, whether public signals are fully informative or not is random. This is why the number of observations varies across sample. * $p < 0.10$, *** $p < 0.01$.

Table 3: Influence of the MPCR on Contributions in the Baseline VCM and Private Signal treatments

	GSU		GATE		GMU		Monash	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
MPCR type (0.25, 0.55, 0.95)	3.218		9.720**		9.174**		2.75	
	(4.897)		(3.890)		(4.584)		(4.101)	
Round number (1 to 8)	-0.263***	-0.336***	-0.186***	-0.264***	-0.200***	-0.250***	-0.242***	-0.308***
	(0.046)	(0.073)	(0.046)	(0.066)	(0.049)	(0.077)	(0.048)	(0.080)
Private signal	2.054	-2.96	-3.594*	-0.202	2.078	-0.594	2.151	1.387
	(3.845)	(3.541)	(2.090)	(4.124)	(2.599)	(2.565)	(2.783)	(3.749)
True MPCR	-8.087		-28.84**		-28.10*		-3.812	
	(17.150)		(13.750)		(16.670)		(15.320)	
Uncertainty	-1.334		0.484		-0.0383		2.351***	
	(0.933)		(0.750)		(0.930)		(0.777)	
Uncertainty X Round number	0.0784**	0.142*	-0.0312	-0.0294	0.0002	-0.0414	-0.0216	-0.032
	(0.034)	(0.075)	(0.026)	(0.081)	(0.035)	(0.088)	(0.030)	(0.095)
True MPCR X Private signal	1.961	8.883**	8.771**	1.537	4.261	5.197	0.191	3.009
	(5.463)	(4.102)	(3.905)	(4.776)	(4.896)	(4.305)	(5.101)	(5.760)
Others' contributions (t - 1)	-0.0443	-0.0659**	0.118***	-0.0197	0.0615**	-0.0443**	0.117***	0.0395
	(0.028)	(0.028)	(0.036)	(0.045)	(0.029)	(0.018)	(0.035)	(0.026)
Others' contrib. (t - 1) X Unc.	0.0849***	0.0302	-0.0371	-0.00527	0.0308	0.000387	-0.0759**	-0.0727**
	(0.032)	(0.037)	(0.032)	(0.050)	(0.033)	(0.032)	(0.032)	(0.034)
Order	0.604		-0.12		-0.294		0.793**	
	(0.373)		(0.342)		(0.394)		(0.343)	
Period (1 to 4)	1.164***		1.429***		1.511***		1.836***	
	(0.360)		(0.335)		(0.268)		(0.357)	
At least 1 signal > True MPCR	-0.34	-0.478	-0.188	0.0469	0.0467	-0.652	-0.802**	-0.583
	(0.292)	(0.721)	(0.389)	(0.663)	(0.459)	(0.457)	(0.378)	(0.646)
At least 1 signal < True MPCR	-0.0233	0.154	0.13	0.667	0.282	0.602	-0.445	-0.362
	(0.269)	(0.515)	(0.296)	(0.584)	(0.494)	(0.545)	(0.324)	(0.466)
Constant	1.251		-1.839***		-1.571*		-3.792***	
	(0.899)		(0.712)		(0.859)		(0.738)	
Number of observations	2,016	2,016	2,016	2,016	2,016	2,016	2,016	2,016
R-squared	0.287	0.595	0.416	0.683	0.329	0.661	0.419	0.648
Number of subjects	96	96	96	96	96	96	96	96

Notes: The models report estimates from linear models with standard errors clustered both at the group and individual levels. The data only includes observations from the *Baseline VCM* treatment and from groups within the *Private Signals* treatments (both *Thin* and *Thick*). Variable "Private signal" refers to the private signal received, and it is equal to the true MPCR in the *Baseline VCM* treatment. Dummy variable "At least 1 signal > True MPCR" equals one when at least one group member received a private signal greater than the true MPCR. Dummy variable "At least 1 signal < True MPCR" equals one when at least one group member received a private signal lower than the true MPCR.

Table 4: Average Contributions in the Baseline VCM and Private Signal treatments

Location	*Baseline VCM*	*Private Signal* Treatments	p-value
	Avg. individual contribution	Avg. individual contribution	
GSU	4.267	4.965	0.054
	(1.698)	(1.667)	
	[32]	[64]	
GATE	4.414	3.95	0.219
	(1.927)	(1.629)	
	[32]	[64]	
GMU	4.544	5.145	0.149
	(1.855)	(1.930)	
	[32]	[64]	
Monash	4.401	4.587	0.621
	(1.746)	(1.735)	
	[32]	[64]	

Notes: Table 4 reports averaged individual contributions across baseline and private signals treatments in our four samples. Standard deviations are in parentheses, and the number of subjects are in square brackets. The last column reports p-values from two-sided t-tests.

Table 5: Replication Table

Power=0.5				
	Successful	Failed		
	Original Study	Replication=1	Replication=2	Replication=3
π		PSP		
0.01	0.05	0.02	0.01	0.01
0.05	0.21	0.12	0.06	0.03
0.10	0.36	0.22	0.12	0.07
0.15	0.47	0.31	0.18	0.10
0.20	**0.56**	0.39	0.24	0.14
0.25	**0.63**	0.46	0.30	0.17
0.30	**0.68**	**0.52**	0.35	0.21
0.35	**0.73**	**0.58**	0.40	0.25
0.40	**0.77**	**0.63**	0.46	0.30
0.45	**0.81**	**0.67**	**0.51**	0.34
0.50	**0.83**	**0.72**	**0.56**	0.39
0.55	**0.86**	**0.76**	**0.61**	0.44

Notes: Table 5 reports the PSP for different priors π after one statistically significant original study, and three subsequent failed replications. The PSP for the successful study is calculated as the PSP from obtaining at least one successful study out of four total studies. The subsequent PSP for failed replications update the PSP for additional failed replications (out of four total studies). We marked in bold PSPs above 50%, that is, cases in which a Bayesian observer believes that it is more likely than not that the significant result is real.

A Online Appendix A

A.1 Appendix Figures and Tables

Table A1: Average Characteristics of the Participants, by Location

	GSU	GATE	GMU	Monash
	(1)	(2)	(3)	(4)
Nb participants	160	160	160	160
Mean age	19.83	21.42***	23.09***	21.63***
S.D.	(1.58)	(1.99)	(3.52)	(3.34)
Mean % of females	0.61	0.54***	0.39***	0.46***
S.D.	(0.49)	(0.50)	(0.49)	(0.50)

Notes: The Table reports average statistics and the results of Mann-Whitney tests (for age) and proportion tests (for gender) comparing each sample to the original GSU sample. S.D. for standard deviations. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Average Round Contributions as Percentage of Endowment, by Treatment and Location when MPCR=0.25

GATE									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	21.9	14.4	9.7	10.3	10.9	9.4	7.5	7.5	11.4
Private Thin	20	13	9.4	9.1	9.7	5.9	6.6	4.4	9.8
Private Thick	23.1	12.8	17.2	7.8	8.8	12.2	10.3	5.6	12.2
Public Thin	27.2	12.8	5	11.7	9.4	5.3	4.8	3.1	9.9
Public Thick	22.8	15.6	15	9.4	7.8	8.4	9.1	6.6	11.8
Total	23	13.7	11.2	9.7	9.3	8.2	7.7	5.4	11
Baseline - Private Thin	1.9	1.4	0.3	1.2	1.2	3.5	0.9	3.1	1.6 ns
Baseline - Private Thick	-1.2	1.6	-7.5	2.5	2.1	-2.8	-2.8	1.9	-0.8*
Baseline - Public Thin	-5.3	1.6	4.7	-1.4	1.5	4.1	2.7	4.4	1.5 ns
Baseline - Public Thick	-0.9	-1.2	-5.3	0.9	3.1	1	-1.6	0.9	-0.4 ns

GMU									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	32.2	22.8	25	16.6	10.9	16.6	8.4	9.4	17.7
Private Thin	29.1	39.4	34.8	26	21.2	19.1	23.8	15.4	26.1
Private Thick	33.4	28.1	29.4	25.6	20.5	20.4	19.3	16.8	24.2
Public Thin	32.4	31.6	28.8	25.8	21.5	17.7	19.1	10.3	23.4
Public Thick	35.9	40.3	38.4	26.6	24.4	19.7	29.1	18.4	29.1
Total	32.6	32.4	31.3	24.1	19.7	18.7	19.9	14.1	24.1
Baseline - Private Thin	3.1	-16.6	-9.8	-9.4	-10.3	-2.5	-15.4	-6	-8.4***
Baseline - Private Thick	-1.2	-5.3	-4.4	-9	-9.6	-3.8	-10.9	-7.4	-6.5***
Baseline - Public Thin	-0.2	-8.8	-3.8	-9.2	-10.6	-1.1	-10.7	-0.9	-5.7*
Baseline - Public Thick	-3.7	-17.5	-13.4	-10	-13.5	-3.1	-20.7	-9	-11.4***

MONASH									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	20	10.6	10.9	5.6	4.7	6.2	8.4	3.4	8.8
Private Thin	38.1	26.6	27.5	19.7	17.8	12.8	9.4	11.2	20.4
Private Thick	31.2	25.6	18.8	12.5	17.8	10	6.9	12.5	16.9
Public Thin	23.4	14.1	9.1	8.1	13.4	10.9	10.6	9.7	12.4
Public Thick	30.3	24.1	16.6	11.9	16.2	10.6	10	6.9	15.8
Total	28.6	20.2	16.6	11.6	14	10.1	9.1	8.8	14.9
Baseline - Private Thin	-18.1	-16	-16.6	-14.1	-13.1	-6.6	-1	-7.8	-11.6***
Baseline - Private Thick	-11.2	-15	-7.9	-6.9	-13.1	-3.8	1.5	-9.1	-8.1***
Baseline - Public Thin	-3.4	-3.5	1.8	-2.5	-8.7	-4.7	-2.2	-6.3	-3.6***
Baseline - Public Thick	-10.3	-13.5	-5.7	-6.3	-11.5	-4.4	-1.6	-3.5	-7***

Table A3: Average Round Contributions as Percentage of Endowment, by Treatment and Location when MPCR = 0.55

GATE									
	Round								
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	53.8	55.3	51.7	49.4	50.6	40.5	41.2	36.6	47.4
Private Thin	40.3	38.4	42	39.7	39.2	37.3	23.6	15.5	34.5
Private Thick	39.4	40	42.5	48.4	38.8	35.6	23.8	22.3	36.4
Public Thin	50.9	49.1	46.6	47.8	32.8	32.7	33.1	35.6	41.1
Public Thick	44.4	45.3	39.1	38.1	27.8	27.5	26.6	25.9	34.3
Total	45.8	45.6	44.4	44.7	37.8	34.7	29.7	27.2	38.7
Baseline - Private Thin	13.5	16.9	9.7	9.7	11.4	3.2	17.6	21.1	12.9***
Baseline - Private Thick	14.4	15.3	9.2	1	11.8	4.9	17.4	14.3	11***
Baseline - Public Thin	2.9	6.2	5.1	1.6	17.8	7.8	8.1	1	6.3*
Baseline - Public Thick	9.4	10	12.6	11.3	22.8	13	14.6	10.7	13.1***

GMU									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	55.3	55	52.5	42.5	42.5	48.4	39.4	33.1	46.1
Private Thin	59.3	72.5	69.4	63	57.7	61.6	51.7	50.6	60.7
Private Thick	49.4	56.6	54.4	55.6	55.6	52.2	45	40.9	51.2
Public Thin	60.2	64.3	56.2	56.5	56.8	53.8	46.5	46.2	55
Public Thick	65.9	62.2	59.4	60.3	58.4	52.8	43.4	44.4	55.9
Total	58	62.1	58.4	55.6	54.2	53.8	45.2	43	53.8
Baseline - Private Thin	-4	-17.5	-16.9	-20.5	-15.2	-13.2	-12.3	-17.5	-14.6***
Baseline - Private Thick	5.9	-1.6	-1.9	-13.1	-13.1	-3.8	-5.6	-7.8	-5.1 ns
Baseline - Public Thin	-4.9	-9.3	-3.7	-14	-14.3	-5.4	-7.1	-13.1	-8.9***
Baseline - Public Thick	-10.6	-7.2	-6.9	-17.8	-15.9	-4.4	-4	-11.3	-9.8***

MONASH									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	60	65.3	60	45.3	41.2	36.9	33.4	26.9	46.1
Private Thin	59.4	58.1	53.8	49.1	48.8	44.4	34.4	31.6	47.4
Private Thick	39.4	42.5	41.2	38.8	34.1	30.3	33.4	30.9	36.3
Public Thin	48.1	58.8	55.6	40.3	34.4	34.4	34.1	27.2	41.6
Public Thick	53.4	51.9	53.1	49.4	53.8	48.1	39.4	35.6	48.1
Total	52.1	55.3	52.8	44.6	42.4	38.8	34.9	30.4	43.9
Baseline - Private Thin	0.6	7.2	6.2	-3.8	-7.6	-7.5	-1	-4.7	-1.3 ns
Baseline - Private Thick	20.6	22.8	18.8	6.5	7.1	6.6	0	-4	9.8***
Baseline - Public Thin	11.9	6.5	4.4	5	6.8	2.5	-0.7	-0.3	4.5 ns
Baseline - Public Thick	6.6	13.4	6.9	-4.1	-12.6	-11.2	-6	-8.7	-2 ns

Table A4: Average Round Contributions as Percentage of Endowment, by Treatment and Location when MPCR = 0.95

GATE									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	73.1	77.5	74.7	78.4	72.8	69.7	60	57.2	70.4
Private Thin	71.9	74.7	71.6	70.3	62.3	60	58.4	57.2	65.8
Private Thick	69.4	77.5	71.2	70	58.8	60.6	58.4	56.2	65.3
Public Thin	73.1	74.7	65.3	61.9	66.9	61.2	59.4	51.2	64.2
Public Thick	70.3	63.1	69.7	61.9	62.5	64.7	52.5	50.6	61.9
Total	71.6	73.5	70.5	68.5	64.7	63.2	57.8	54.5	65.5
Baseline - Private Thin	1.2	2.8	3.1	8.1	10.5	9.7	1.6	0	4.6 ns
Baseline - Private Thick	3.7	0	3.5	8.4	14	9.1	1.6	1	5.1 ns
Baseline - Public Thin	0	2.8	9.4	16.5	5.9	8.5	0.6	6	6.2 ns
Baseline - Public Thick	2.8	14.4	5	16.5	10.3	5	7.5	6.6	8.5*

GMU									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	75.6	70.6	69.1	65.6	64.1	57.8	64.1	70	67.1
Private Thin	78.7	75.6	74.5	69.1	68.8	67.3	65.7	69.5	71.1
Private Thick	75.9	77.8	77.2	79.4	76.2	74.4	70.6	56.9	73.6
Public Thin	85	82.2	79.7	78.1	72.2	69.7	66.6	62.8	74.5
Public Thick	78.1	76.6	77.8	75.6	70	69.4	66.9	67.5	72.7
Total	78.7	76.6	75.6	73.6	70.3	67.7	66.8	65.3	71.8
Baseline - Private Thin	-3.1	-5	-5.4	-3.5	-4.7	-9.5	-1.6	0.5	-4 ns
Baseline - Private Thick	-0.3	-7.2	-8.1	-13.8	-12.1	-16.6	-6.5	13.1	-6.5 **
Baseline - Public Thin	-9.4	-11.6	-10.6	-12.5	-8.1	-11.9	-2.5	7.2	-7.4 ns
Baseline - Public Thick	-2.5	-6	-8.7	-10	-5.9	-11.6	-2.8	2.5	-5.6 ns

MONASH									
					Round				
Treatment	1	2	3	4	5	6	7	8	Total
Baseline VCM	68.8	75	71.2	77.5	69.7	61.9	56.2	60.9	67.7
Private Thin	79.7	78.4	81.6	76.6	72.5	67.5	60	50.3	70.8
Private Thick	80.6	77.5	74.1	73.8	73.4	66.2	66.6	70.3	72.8
Public Thin	80.9	75	69.7	64.7	67.5	70.6	61.9	52.5	67.9
Public Thick	83.4	80.3	80	71.9	71.9	66.6	65	56.6	72
Total	78.7	77.2	75.3	72.9	71	66.6	61.9	58.1	70.2
Baseline - Private Thin	-10.9	-3.4	-10.4	0.9	-2.8	-5.6	-3.8	10.6	-3.1 ns
Baseline - Private Thick	-11.8	-2.5	-2.9	3.7	-3.7	-4.3	-10.4	-9.4	-5.1 ns
Baseline - Public Thin	-12.1	0	1.9	12.8	2.2	-8.7	-5.7	8.4	-0.2 ns
Baseline - Public Thick	-14.6	-5.3	-8.8	5.6	-2.2	-4.7	-8.8	4.3	-4.3*

Table A5: Effect of Public Signals' Informativeness on Cooperation

	GSU	GATE	GMU	Monash
	(1)	(2)	(3)	(4)
	Contribution	Contribution	Contribution	Contribution
Round number (1 to 8)	-0.383***	-0.135	-0.316***	-0.305**
	(0.129)	(0.117)	(0.110)	(0.120)
True MPCR	3.657***	10.06***	5.352***	8.003***
	(1.060)	(1.064)	(1.232)	(1.098)
N. MPCRs compatible with signal	-0.233	-0.935	0.741	6.563
	(1.565)	(2.456)	(1.461)	(4.798)
N. compatible MPCRs squared	-0.055	0.316	-0.122	-1.145
	(0.202)	(0.384)	(0.202)	(0.908)
Only one possible MPCR	0.787	-1.246	0.725	2.736
	(0.967)	(1.117)	(0.961)	(2.242)
True MPCR X n. compatible MPCRs	1.081*	-1.411**	0.0528	-0.238
	(0.586)	(0.567)	(0.639)	(0.715)
True MPCR X Round	0.176	-0.101	0.112	0.0963
	(0.164)	(0.159)	(0.188)	(0.222)
Round X n. possible MPCRs	0.105	-0.072	-0.0234	0.0782
	(0.077)	(0.072)	(0.061)	(0.078)
Round X True MPCR X n. possible MPCRs	-0.082	0.0866	0.0609	-0.142
	(0.091)	(0.086)	(0.102)	(0.143)
Number of observations	3,072	3,072	3,072	3,072
R-squared	0.559	0.672	0.572	0.636

Notes: The table reports estimates from a lineal model with standard errors clustered both at the group and individual level. The model in all samples include the *Baseline VCM* treatment and all observations from Public Signals treatments (both *Thin* and *Thick*). "True MPCR" identifies the true MPCR for the round. "Number of possible MPCRs compatible with all signals" counts the number of values that are compatible with the true MPCR given the public signals. The dummy "Only one possible MPCR (0=no; 1=yes)" takes value 0 when the public signals do not uniquely identify the true MPCR, and 1 when they do (and in all observations in the *Baseline VCM*). Robust standard errors are in parentheses (for models 2 and 4). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Determinants of Contributions between MPCR (Tobit)

	GSU	GATE	GMU	Monash
	(1)	(2)	(3)	(4)
MPCR type	24.67***	21.09**	22.88***	3.84
	(7.723)	(9.152)	(8.077)	(13.530)
Round number (1 to 8)	-0.496***	-0.541***	-0.463***	-0.739***
	(0.071)	(0.091)	(0.081)	(0.108)
Private signal	-6.412*	-6.189	-1.494	8.062
	(3.396)	(4.245)	(2.838)	(5.031)
True MPCR	-68.11***	-65.75**	-64.15***	1.482
	(22.280)	(29.630)	(23.340)	(38.720)
Uncertainty	3.368	0.762	0.0232	0.189
	(2.562)	(3.103)	(2.723)	(3.520)
Uncertainty X Round number	0.195**	-0.0451	0.00789	0.0953
	(0.086)	(0.110)	(0.098)	(0.128)
True MPCR X Private signal	14.69***	8.345	7.631*	-3.619
	(4.604)	(5.805)	(4.366)	(7.162)
Others' contributions (t - 1)	-0.108***	-0.0259	-0.0861**	0.0765*
	(0.035)	(0.047)	(0.038)	(0.044)
Others' contribution (t - 1) X Uncertainty	0.0348	-0.0125	-0.00317	-0.140**
	(0.043)	(0.055)	(0.048)	(0.054)
Order	0.686	-10.68***	-2.243	5.194*
	(1.342)	(2.732)	(2.181)	(2.845)
Period (1 to 4)	2.790***	9.683***	3.229***	6.838***
	(0.853)	(1.261)	(0.900)	(1.276)
At least 1 member: signal> True MPCR	-1.105*	-0.918	-0.902*	-0.185
	(0.661)	(0.702)	(0.524)	(0.848)
At least 1 member: signal< True MPCR	0.242	0.745	0.947	-0.734
	(0.601)	(0.757)	(0.594)	(0.900)
Constant	-10.98	-0.129	-1.212	-20.41**
	(6.956)	(5.979)	(7.506)	(10.390)
Number of observations	2,016	2,016	2,016	2,016
R-squared				
Number of subjects	96	96	96	96

Notes: All models report estimates from Tobit models. The data only includes observations from the *Baseline VCM* treatment and from groups within the *Private Signals* treatments (both *Thin* and *Thick*). Variable "Private signal" refers to the private signal received, and it is equal to the true MPCR in the *Baseline VCM* treatment. Dummy variable "At least 1 signal > True MPCR" equals one when at least one group member received a private signal greater than the true MPCR. Dummy variable "At least 1 signal < True MPCR" equals one when at least one group member received a private signal lower than the true MPCR. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.