HAL
open science

# A New Mechanism to Alleviate the Crises of Confidence in Science With An Application to the Public Goods GameA Review

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

## To cite this version:

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 GameA Review. 2020. halshs-02512932

**HAL Id: halshs-02512932**

**https://shs.hal.science/halshs-02512932v1**

Preprint submitted on 20 Mar 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GATE LYON SAINT-ÉTIENNE
UMR 5824

93, chemin des Mouilles
69130 Ecully - France

Maison de l'Université, Bâtiment B
10, rue Tréfilerie
42023 Saint-Etienne cedex 02 - France

http://www.gate.cnrs.fr
gate@gate.cnrs.fr

WP 2011 – March 2020

# 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, Marie Claire Villeval

**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' incentives are not sufficient to encourage replications. We analyze a novel mechanism promoting replications through beneficial gains between scholars and editors. We highlight the tradeoffs 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 but largely unexplored feature in the literature.

**Keywords:**

Replication, science, public goods, uncertainty, experiment

**JEL codes:**
A11, C18, C92, D82

Working Paper

# 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[‖]

March 19, 2020

**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' incentives are not sufficient to encourage replications. We analyze a novel mechanism promoting replications through beneficial gains between scholars and editors. We highlight the tradeoffs 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 but largely unexplored feature 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, Anna Dreber, Lata Gangadharan, Håkan Holm, David Jimenez-Gomez, Johanna Mollerstrom, Fatemeh Momeni, John Nye, Giovanni Ponti, Adam Sanjurjo, 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). 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 social and moral 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. Within this movement, and in addition to the development of laboratory experiments, economists now view domains as distinct as classrooms, boardrooms, open-air markets, and automobile plants as fertile grounds to explore their economic hypotheses (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 on the topic, can significantly accelerate the advancement of economic science. Nonetheless, the field of economics, in its current state, presents few strong incentives to replicate. Once a study has been published, the original investigators have little incentive to replicate their own findings. The reason is that the returns from replicating published work are generally low.[1] This is problematic, as new and surprising findings may be false positives simply due to the mechanics of statistical inference (*e.g.*, Coffman and Niederle (2015); Dreber et al. (2015); Coffman et al. (2017); Maniadis et al. (2017)). Similarly, new and surprising studies may suffer from low power or weak initial support, and thus may be dismissed though they point toward an economic association that is ultimately true. Because novel results attract attention, and are generally sought by academic journals much more than replication studies, an important question of incen-

[1] As pointed out by Coffman and Niederle (2015), attempting to replicate someone else's work may even generate animosity, further reducing the incentives to replicate.

tives arises.[2]

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, they commit to never 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 at the AEA experimental registry and referenced back in the first working paper.[3] This guarantees that all replications, whether successful or unsuccessful, are properly recognized. 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 taps into the core of the incentive problem that editors, original authors, and replicators face. We highlight in a model that within this three-player market, certain features must be in place for the mechanism to be incentive compatible. For example, within our model, if at least one of two conditions is present, this mechanism might be chosen over the *status quo* publishing approach. One condition is that journal editors prefer empirical results that have been independently replicated, *ceteris paribus*. A second condition is that scientists value reporting the truth, as they prefer their own published work to be correct. This could arise either from the original authors suffering disutility from publishing a study that is later found to be a false positive (not replicable) or from scientists who prefer that public

[2]While this paper focuses on the close replication of an existing experimental design, other types of replications are rare as well, such as obtaining published datasets to replicate the results, or investigating a research question using a different design and setting. For instance, Hamermesh (2007) surveyed authors of 139 empirical studies published between 2002 and 2004 in *Industrial and Labor Relations Review* and the *Journal of Human Resources*, both journals with open data access policies. He found that the mean number of requests for data in each of these two specialized journals was just one, and 60.5% of the authors of these papers never received a request to share their data. Hamermesh (2007) conducted a similar survey with authors who published in the *American Economic Review* between 1999 and 2000, and found that the median request for data was three, with 22% of authors never receiving a request. By contrast, there now seems to be a burgeoning demand for replication studies (see, for example, Benjamin et al. (2018); Camerer et al. (2018)).

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

resources are not used for false positives (*i.e.*, policies put in place that do not work). Our simple theoretical framework highlights how these and other conditions lead to the approach being incentive compatible. 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 stronger publications that would otherwise be hard to attain.

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 effect sizes, the mechanism under investigation is instead geared towards using a Bayesian model of independence to update priors, as well as accounting for potentially important differences across experimental environments.

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 (2019) 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 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 the replication mechanism, the working paper was distributed online, but never published. 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 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.[4] Yet, when considering the basic result of whether Knightian uncertainty facilitates overall cooperation, the original results do not replicate using a stringent 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 surpris-

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

ing 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" 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 raised 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.[5] This approach, while desirable, does not fully address the fact that replication studies generally carry low returns in terms of academic prestige (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.[6]

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 that the replication effort is ultimately published.[7] Further, Maniadis et al. (2015) sug-

[5]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)*".

[6]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.

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

gest 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. 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. (2019)). Furthermore, the mechanism we analyze is particularly well-suited to handle studies whose surprising results are very likely to generate low priors.[8]

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, 2018). 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 (Abadie, 2018; Kessler and Meier, 2014), 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 be directly used to confirm a null hypothesis, the Post-Study Probability derived from a series of independent failed replications nevertheless 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,

[8] See also Camerer et al. (2016, 2018)

one of the most studied decision-making environments in the field of experimental economics (see, (Andreoni, 1995) for an early contribution). 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 (Forthcoming).

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.[9] 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.

# 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). Finally, it shows why the mechanism under investigation may be particularly well-suited to original studies likely to suffer from low priors.

## 2.1 The Importance of Replications

Maniadis et al. (2015) propose a simple Bayesian framework to evaluate how novel results should move scholars' priors.[10] Let us start with the simplest case of updating after ob-

[9] 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.

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

serving results from one study. Let $\pi$ be the prior that a given scholar has about a given scientific relationship. Call $\alpha$ 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 $\pi$, 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 $\pi$). So, for instance, if a given scholar believed that a certain scientific result had a 1% chance of being true at $\alpha = 5\%$ level and power $(1-\beta) = 80\%$, after observing one study confirming that result, he would update his priors to 13.9%.

Even more dramatically, a scholar holding priors of 10% would update the post-study probability to 64%. This exercise highlights how volatile low priors are when they only depend on evidence provided by a single study. Figure 1 shows how Bayesian scholars holding initial priors $\pi = 1\%$ and 10% should update their posteriors based on subsequent failed replications. With one additional failed replication following a significant initial result, the PSP given initial priors $\pi = 1\%$ would fall from 13.9% (the PSP after first significant result) to about 3%, and to 0.07% with two failed replications. Assuming initial priors $\pi = 10\%$, one failed replication would lower the PSP from 64% to 27.2%, while two would further reduce the PSP to 7.3%. It can be easily seen that with three or more failed replications, the post-study probability converges toward zero, regardless of the initial priors.

Similarly, just a few successful replications allow robust convergence of PSP at above 80%, regardless of initial priors. Figure 2 shows how the PSP (assuming $\pi = 1\%$) varies based on the number of successful replications out of five and out of ten total replication attempts.

The message from these illustrations is clear: a few independent replications allow for a wide range of beliefs to converge. We suspect that this is why Fischer's (1935) original tripod included replications as one key feature. As aforementioned, however, there are few professional incentives for a wider and 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 4 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 data, but agree that they will never 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 coauthorship 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.[11] 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 signal credibly the paternity of the original research idea, and, as explained below, it provides a binding commitment device for original authors and other scholars alike that increases the credibility 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. 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.[12] 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.[13]

[11] 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.

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

[13] 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

*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 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 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.

Two outcomes are possible: the replication is successful or unsuccessful. If the replication is successful, then the PSP that referees would rationally assign to the results would not impact the paper's reception. This is because successful replications induce referees' posterior beliefs to converge, regardless of their priors.[14] For any given journal, a success-

provides a credible signal about the total number of replications that will be conducted. This is critical to avoid unethical behavior, e.g., including in the final paper only successful replications.

[14]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

fully 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. As we gain a firmer understanding of the value of null results, we foresee such robust null results increasing in import.

Second, beyond the benefits provided to all authors who care about the robustness of their results, the mechanism described has a particularly pronounced value for junior scholars. The authors of a novel study may choose to replicate their own work themselves to increase the chance of a successful publication, rather than resorting to this replication mechanism. This may be relatively easy for senior scholars, who likely have easier access to financial resources, but less so for junior scholars. Instead, the mechanism we investigate here externalizes the cost of replications. Moreover, journals may have a taste for novelty that may, all else equal, increase the probability of a novel and surprising study being published. In such a circumstance, the authors of a novel study might prefer attempting to publish it immediately rather than seeking replication. While both junior and senior scholars may attempt to publish a novel study on its own, in practice, senior scholars may be more likely to succeed in this process. There are a number of reasons for this. Namely, they are more likely to have an established reputation for rigorous scientific conduct, relatively lower pressure from the publish-or-perish culture, and perhaps a relatively stronger influence on the editorial process. In contrast, junior scholars may face greater obstacles, as their reputations are not yet well established. In this manner, this mechanism can help junior scholars establish their reputations by submitting replicated papers 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 affect their choice of trying to publish their work immediately or after replication. Thus, there is a trade-off.

In investigating this mechanism, we anticipate that not only the original authors, but also replicators and editors, are involved in the trade-offs to be made when deciding to use the mechanism to replicate an original study. In the next section, we sketch a simple model to highlight the important issues of when and where the mechanism we study might be particularly useful.

---

studies would move posteriors more than a single study, even in presence of the bias $u$.

## 3 Theoretical Framework: The Replication Dilemma

In this section, we outline a simple model capturing the trade-offs involved in seeking independent replications of one's own work before submitting it to journals. While the framework can encompass different ways of generating replications, we directly refer to the method proposed by Butera and List (2017) (BL, hereafter). The goal is twofold: first, we wish to highlight the relevant parameters and preferences influencing the decision to replicate a novel study. Second, we aim to show conditions wherein scholars have an incentive to replicate. Overall, the analysis points to the critical role of journal editors in encouraging replications.

In its most general form, the "replication dilemma" can be thought of as a strategic game composed of three players: (i) the authors of an original study, (ii) the journal editors, and (iii) other scholars who may or may not replicate their colleagues' work ("replicators"). In this general setting, authors optimally choose either to replicate their original work before submitting to journals, or to publish without first seeking replications. Editors' equilibrium preferences for replicated work are determined by their individual preferences for papers' novelty and scientific robustness, and by the competition among editors for valuable articles. Finally, other scholars decide how to allocate their time and resources between generating original work and joining other authors in their replication efforts, based on the relative returns of these investments. Given that the focus here is on the incentives for the original authors to replicate, we treat editors and replicators' equilibrium behavior as exogenous. In particular, we assume that there is an infinite supply of replicators.[15] We begin by describing the role of editors. Then, we describe the preferences of the original authors. Finally, we characterize the decision problem faced by the original authors, and discuss the conditions that induce replication.

Editors in our model are a singleton with exogenous preferences, and are characterized by two parameters. The first parameter $E \in (0, 1)$ captures editors' preferences over publishing novel results that have been replicated (either successfully or not). When $E = 0.5$, editors are all else equal indifferent between publishing replicated and nonreplicated papers. When $E = 1$, editors would only publish work that has been replicated, either successfully or unsuccessfully.[16] The second parameter $\phi \in (0, 1)$ captures editors' preferences for suc-

[15] This implies that in our model, replicators face no opportunity cost for their time.

[16] Note that while we set $E \in (0, 1)$, a lower bound of 0.5 is generally more plausible since $E < 0.5$ would imply that, all else equal, editors would strictly prefer to publish results that have not been replicated. This seems unrealistic for most journals and situations. Alternatively, one could imagine a counterfactual

cessful replications, that is, $\phi$ captures by how much editors discount failed replications. If $\phi = 1$, then editors would be equally likely to accept papers reporting on either successful or unsuccessful replications.

We now turn to the preferences of the original authors. First, we assume that scholars may care about finding and disseminating scientifically valid results, as opposed to false positives (or false negatives). We summarize these preferences through parameter $\tau \in (0, 1)$. As $\tau$ increases, authors place greater value on replications, both successful and unsuccessful, and both solicited by them or conducted by other scholars after their initial publication. The reason is that replications allow research inquiries to converge toward scientific truths. Authors also value their academic reputations $R \in (0, 1)$, and face the replication problem at different levels of seniority $S \in (0, 1)$: a lower $S$ corresponds to more junior authors, and $S = 1$ implies that an author has been granted tenure. Authors have preferences over the quality of the journal that publishes their work. For simplicity, we assume that the quality of the journal is captured through a simple numeraire $J$. Authors also have a time discount factor $\delta$, and their patience increases as they age, due to tenure requirements. That is, they discount future publications at rate $S \cdot \delta$. Finally, scholars have priors about the validity of their results. In particular, we assume that scholars have priors $\pi_R \in (0, 1)$ about the likelihood that their results would successfully replicate.[17]

With this setup, we now describe the decision problem faced by the original authors of a novel study. Figure 3 provides a representation of the extensive form game faced by the original authors. We discuss the replication problem through the lens of a simple two-period model. In the first period, $t = 1$, the authors of a novel study decide whether to publish their work in $t = 1$ without seeking replication, or to wait until $t = 2$, so that replication can be attempted and the paper submitted (using BL). We first discuss the authors' payoffs associated with these two options. We then discuss how changes in three key parameters $(\pi_R, \tau, S)$ affect the replication decision.

Suppose first that the authors decide to submit their original work on their own without seeking to replicate. Suppose further that they are able to publish their paper in $t = 1$. We assume that the value of such publication is equal to $(1 - E) \cdot J \cdot S$. Intuitively, as

situation in which editors of very new or very low-ranked journals try to attract articles, or editors of predatory journals try to convince scholars to pay to publish. In this case, unreplicated work may likely be favored. The reason however is not that these editors dislike replicated work, but simply that replicated papers would not be submitted to their journals in the first place.

[17] We assume that scholars' priors about replicability are independent of whether replications are carried out through the BL mechanism or by other researchers at a later time.

editors' preferences for replicated work increases (*e.g.*, $E$ is large), the lower the returns from non-replicated research. Further, seniority – or experience – $S$ facilitates publishing in better academic outlets: without replications, younger researchers may find it more difficult to publish novel work in highly ranked journals. Authors publish in $t = 1$, but they may also receive additional payoffs in $t = 2$. Payoffs received in $t = 2$ depend on two factors: whether anyone in the future tries to independently replicate the authors' published research, and whether such attempts, if any, are successful.

First, consider the case in which no one else replicates. We assume that the original paper is "ignored" by subsequent literature with probability $p$. If a paper is ignored, then the expected payoffs of the original authors from $t = 1$ perspective take the following form:

$$EV(Solo_{no\ reps}) = (1 - E) \cdot J \cdot S - \tau \cdot \delta \cdot C_\tau \tag{2}$$

The first term of equation 2 captures the benefit of publishing in $t = 1$. The second term instead captures the idea that scholars who place a positive value on the discovery and dissemination of scientifically valid results (*i.e.*, $\tau > 0$), do experience disutility $C_\tau$ when no one replicates existing research. For simplicity, we assume that $C_\tau$ is a numeraire equal to $J$.

Next, consider the case in which other scholars subsequently and independently replicate in $t = 2$ the original paper, with probability $1-p$. With probability $q$ such independent replications will be successful. In this case, the expected value from publishing alone in $t = 1$ and being independently and successfully replicated is:

$$EV(Solo_{succ\ indep\ reps}) = (1 - E) \cdot J \cdot S + \delta \cdot P \cdot (\tau + R) \cdot \pi_R \tag{3}$$

The second term of equation 3 captures the marginal benefits $P$ received in $t = 2$ from having one's work successfully replicated independently ($P = C_\tau = J$). These benefits derive from reputation $R$ and value for science $\tau$, and from the $t = 1$ perspective they have a likelihood of materializing equal to $\pi_R$ (the authors' priors about the likelihood that their research would successfully replicate).

With probability $(1 - q)$ instead, replications will be unsuccessful. In this case, the expected value equals:

$$EV(Solo_{fail\ indep\ reps}) = (1 - E) \cdot J \cdot S + \delta \cdot P \cdot (\tau - R) \cdot (1 - \pi_R) \tag{4}$$

As for equation 3, if $\tau > 0$, then the original authors derive a positive benefit from being replicated, even if unsuccessfully. However, failed replications generate a reputational cost, $P \cdot R$. Such costs and benefits from failed independent replications have a subjective likelihood equal to $(1 - \pi_R)$ from $t = 1$ perspective.

Suppose now that the authors decide to replicate using BL. In this case, their payoff in $t = 1$ will be equal to zero, since they will need to wait for all replications to be completed in $t = 2$ to submit their paper. In $t = 2$, payoffs will depend on whether the original results replicate. From $t = 1$ perspective, the expected value of publishing successfully and unsuccessfully replicated research is:

$$EV(BL_{Success}) = 0 + S \cdot \delta \cdot \pi_R[J \cdot (\tau + R + 1) \cdot E] \tag{5}$$

$$EV(BL_{Fail}) = 0 + S \cdot \delta \cdot (1 - \pi_R)[J \cdot (\tau + R + 1) \cdot \phi E] \tag{6}$$

Equations 5 and 6 show that in $t = 1$, the authors of a novel research receive a payoff of zero. From $t = 1$ perspective, the value of publishing replicated research in $t = 2$ depends on a number of factors. The higher authors discount the future – and the more junior the authors are $(S \cdot \delta)$ – the less appealing replication is, regardless of whether replications were successful. The reason is that junior scholars must publish quickly to secure tenure. Moreover, the value of replicating crucially depends on the weight editors place on replicated work: the higher the value $E$ editors place on replicated work (both successful and unsuccessful), the greater the appeal of replication. However, while editors may highly value replications (*e.g.*, $E$ is large), they may nevertheless be reluctant to publish failed replications (*e.g.*, $\phi$ is small). The appeal of replicating therefore depends on scholars' priors about the likelihood that their original study will replicate–as priors $\pi_R$ increase, the appeal of replication increases since the authors will not expect to be affected by a possibly low $\phi$. Finally, the values that the authors place on their reputation $R$ and on scientifically valid results $\tau$ increase the expected value of replicating, both successfully and unsuccessfully.

We are now in a position to characterize simple comparative statics about the choice to replicate by varying three core parameters of the model: scholars' priors $\pi_R$ about the likelihood that their results replicate; the value scholars place on science, $\tau$; and the level of seniority of scholars, $S$. The details of the exercise are given in section A.1 in Appendix A.

We first consider variations in the priors $\pi_R$. Consider the case in which the original authors had very little confidence in the replicability of their own work. In the limiting case where $\pi_R \simeq 0$, the authors believe that their results will never successfully replicate. In this case, the authors will choose to replicate if and only if the expected value from a collection of failed replications is higher than the expected value from publishing the paper alone. All else equal, $\forall \phi < 1, E_{\pi_R=0} > E_{\pi_R=1}$, that is, the lower bound of editors' tastes for replications that makes replications appealing to scholars is higher for scholars with low priors $\pi_R = 0$ compared to scholars with high priors $\pi_R = 1$. This implies that if $E$ is between $E_{\pi_R=0}$ and $E_{\pi_R=1}$, then authors with higher $\pi_R$ will choose to replicate, while authors with lower $\pi_R$ will choose not to replicate.

We next consider variations in the value $\tau$ authors place on producing and disseminating scientifically valid results. We find that $E_{\tau=0} > E_{\tau>0}$; that is, scholars who place little value on science require a larger lower bound of $E$ to be willing to replicate relative to scholars with $\tau > 0$.

Finally, we consider variations in the level of seniority $S$. Increasing seniority has an ambiguous effect on the lower bound $E_{S>0}$. Suppose first that the original paper will never be replicated; that is, $p = 1$. If scholars place no value on science ($\tau = 0$), then increases in seniority will have no effect on the likelihood of replicating. If instead $\tau > 0$, then the lower bound of $E$ making replications appealing will be higher for seniors than for juniors (*e.g.*, $\frac{\partial E_{S,\tau>0}}{\partial S} > 0$). The reason is that there is no risk of seeing one's paper falsified, and as scholars become more experienced, they become more capable of publishing in highly ranked journals without preliminary replications.

Next, suppose that the original paper will definitely be replicated (*e.g.*, $p = 0$). For successfully replicated papers ($q = 1$), and for any positive values of science $\tau$ and reputation $R$, increases in seniority will reduce the lower bound $E$ necessary to make replications appealing. This is due to the fact that scholars become more patient and know that their work will replicate. Differently, for unsuccessfully replicated papers, whether seniority increases or decreases the lower bound $E$ depends on the relative importance scholars place on science $\tau$ and reputation $R$. If scholars value science more than their own reputations, $\tau > R$, then seniors will require a smaller lower bound of $E$ to replicate compared to juniors. Nevertheless, if scholars value reputation more than science, $\tau < R$, then seniority will increase the lower bound of editors' preferences $E$ necessary to make replications appealing. The reason is that when seniority increases, the reputation drop is compensated by publications in better ranked journals (remember that in $t = 1$ payoffs are $(1 - E) \cdot J \cdot S$

from publishing without replicating).[18]

# 4 Experimental Design and Replication Protocol

We now demonstrate the operation of this mechanism in an experiment on the effects of 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.[19] Second, cooperation decays over time (Andreoni, 1995), a tendency linked to the presence of heterogeneous preferences such as self-interest, altruism, and (sometimes self-serving) conditional cooperation.[20] Third, centralized institutions such as taxation, competition, and voting rules,[21] and decentralized institutions such as communication, moral and monetary sanctioning and rewards[22] 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 approach.

## 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

[18] We implicitly assume that the reputational drop from failed replications is independent of journal quality. One could alternatively argue that the reputational drop from failed replications of a highly ranked publication is much greater than the reputational drop from a failed replication of a relatively minor publication. Such a change will affect decision thresholds but does not alter the general intuition of our model concerning marginal benefit and marginal cost trade-offs.

[19] 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).

[20] 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).

[21] 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).

[22] 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).

Capita Return (MPCR, hereafter) $\theta$ to all group members, regardless of their individual contributions. There is no Knightian uncertainty in this game, as $\theta$ 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 $\theta$, each participant receives a noisy signal, $s_i = \theta + \varepsilon_i$, where $\varepsilon_i$ is distributed according to an unknown distribution, with mean zero and standard deviation $\sigma$. 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:

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

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 $\theta$ 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 $\theta$: If at least two group members receive opposite extreme signals, then $\theta$ 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, partic-

ipants choose how to allocate 10 tokens between a private account and a group account.[23] 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 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 $\pm 0.1$. 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 $\pm 0.2$. 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

[23] 20 tokens are worth U.S. $1.

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 were 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%.[24] Then, the original authors invited coauthors for the second paper through their professional network.[25]

Each replication closely followed the protocol used in the original experiment, including utilizing the same sample size, the same software, and the same instructions.[26] In total, 640 subjects participated in the experiment (160 in the original study and in each replication, equally balanced across treatments).[27] All subjects were students in local universities. Table A1 in the Appendix shows a balance table for gender and age composition. To ac-

[24] 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.

[25] After the three coauthors accepted, 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.

[26] 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.

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

count 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.

## 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.[28]

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.[29] 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

[28] Tables B1, B2, B3 in Appendix provide detailed summary statistics by round and MPCR.

[29] 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*.

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.[30] 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 *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 insignifi-

[30] Note that these are not individual averages.

cance, 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.[31] First, our 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.[32] This

[31] 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.

[32] 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 B4 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).

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.[33]

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 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.[34] 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

[33] 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 $\eta_i$ is an individual fixed effect and $\gamma_g$ is a group fixed effect.

[34] The same holds for our panel tobit specification, see Table B5 in Appendix.

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 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 $\pi$, after our initial (significant) study and after our three (failed) replications.

Table 5 conveys three critical messages. First, small deviations in priors $\pi$ 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 a single 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 mechanism to promote replications within the sciences. The mechanism is simple: upon completion of a study finding significant but surprising results, the authors make the working paper available online, but commit to never submitting it to a journal for publication. Instead, they offer coauthorship for a second, yet to be written paper to scholars willing to independently replicate the study at their own cost. The second paper references the original working paper, includes all preregistered replications, and is submitted to a peer-reviewed journal.

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 suffer from low priors, the successful publication of these studies relies heavily on small variations in the distribution of prior beliefs.[35] 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.

Clearly, this mechanism is only a first step in the direction of promoting a more widespread use of replications in economics, and does not directly address a number of empirical questions.

First, our current model is silent relative to how to choose optimally 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: First, we assumed no bias $u$ in the PSP (neither sympathetic nor antagonistic). Second, 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 this 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

[35] Notice that this reasoning abstracts entirely from the economic relevance of the phenomenon under investigation, meaning that even papers addressing highly compelling problems may still fail to place in top journals due to the simple mechanic of inference. 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).

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.

Second, the mechanism described in this paper is 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 implement 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 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 these weights. These factors would therefore affect the opportunity cost of joining a replication paper.[36]

Third, 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. That said, barriers to entry, especially for young experimental and behavioral economists, already exist and are substantial: laboratory experiments' costs can increase

[36] Another important factor that might encourage junior scholars to adopt a replication mechanism is the reluctance editors may have in asking for more data to junior scholars, knowing how taxing this investment would be for them. As a result, difficult editorial decisions on papers from junior scholars may often tilt towards rejection if asking for major revisions with additional data is perceived by editors as a delicate ask.

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.

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 RCT may simply be infeasible. 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.[37] 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.

[37] One caveat in this case is that differences in the design would make it difficult to exactly compute *ex ante* the PSP.

## References


Abadie, A. (2018): "Statistical Non-Significance in Empirical Economics," Mimeo.

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.

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.

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.

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.

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. (2019): "Crowdsourcing hypothesis tests: Making transparent how design choices shape research results," Psychological Bulletin.

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.

Maniadis, Z., F. Tufano, and J. A. List (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.

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 (Forthcoming): "Cooperation in a risky world," Journal of Public Economic Theory.

Villeval, M. C. (2019): "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.

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 and Tables

*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* $\alpha = 0.05, (1-\beta) = 80\%$*)*

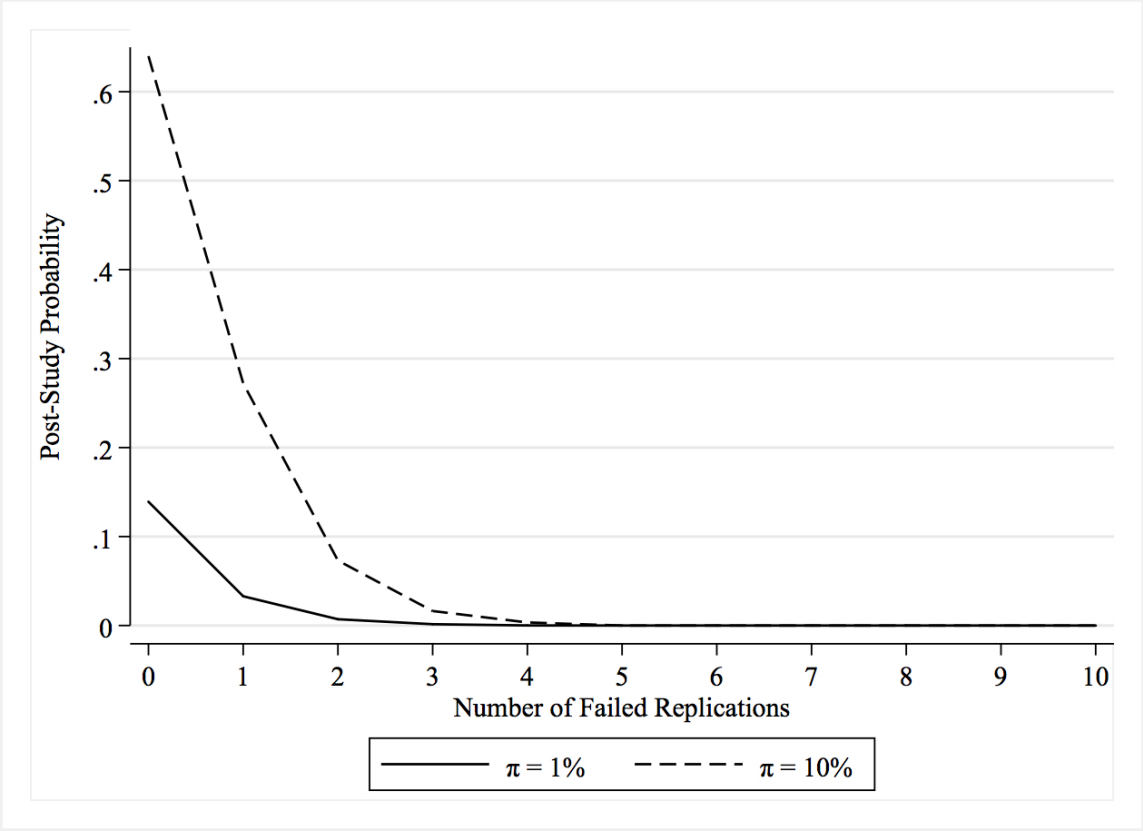


*Figure 2: Post-Study Probability (PSP) of a Given Result Being True as a Function of the Number of Successful Replications (assuming $\pi = 0.01, \alpha = 0.05, (1-\beta) = 80\%$)*

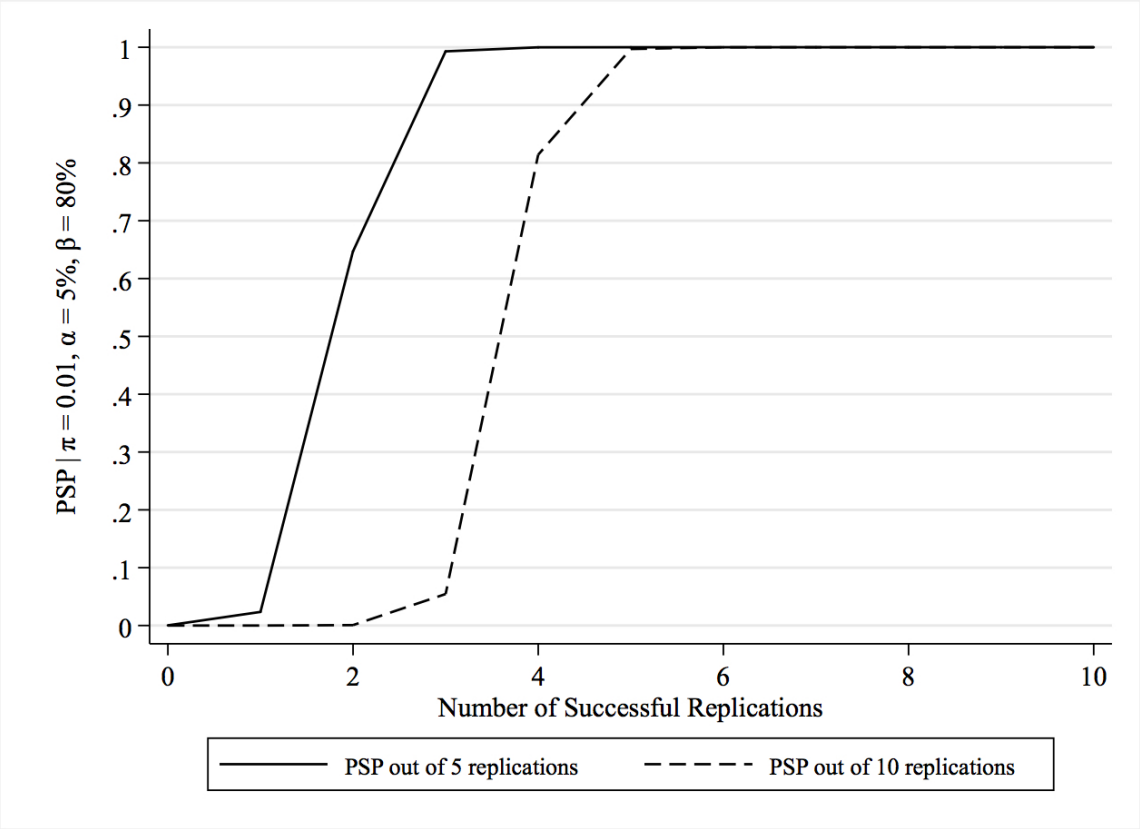


*Figure 3: Decision Problem Faced by the Original Authors of a Novel Study Regarding Replication*

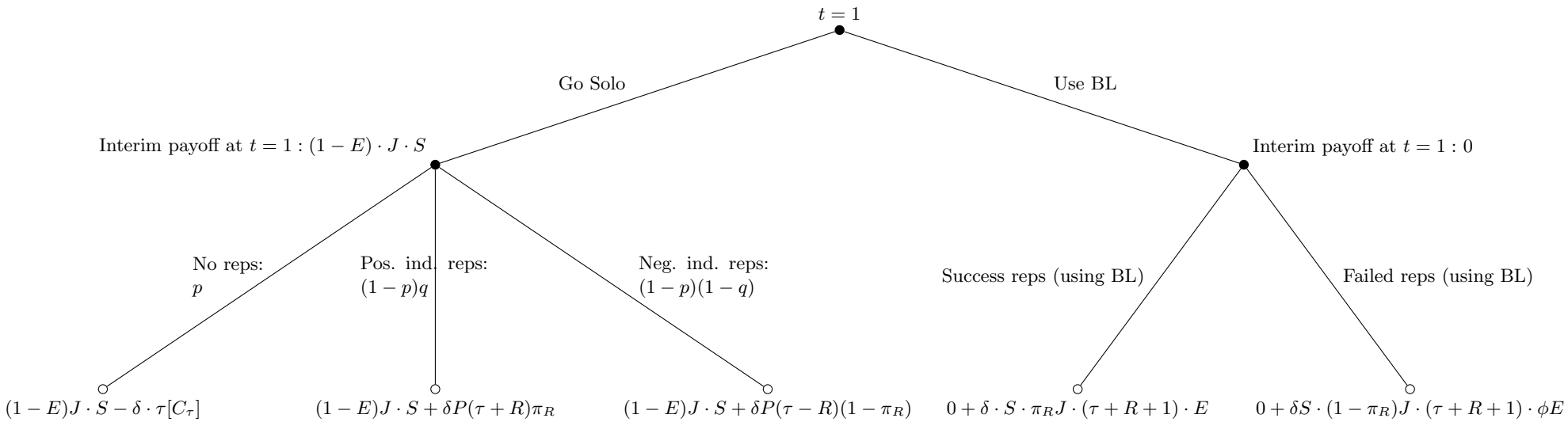
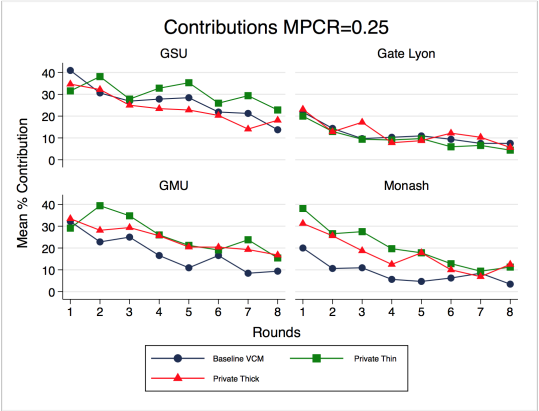
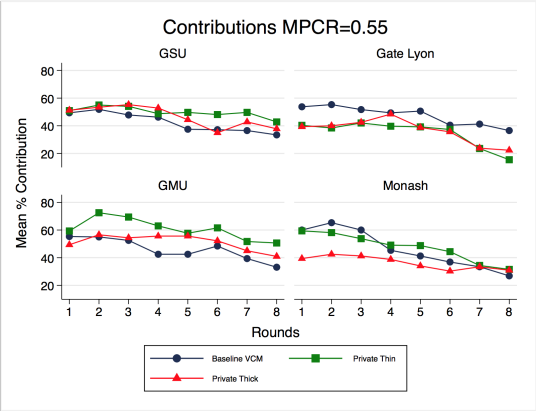


*Figure 4: Average contributions (%) by round, MPCR, sample and treatments*



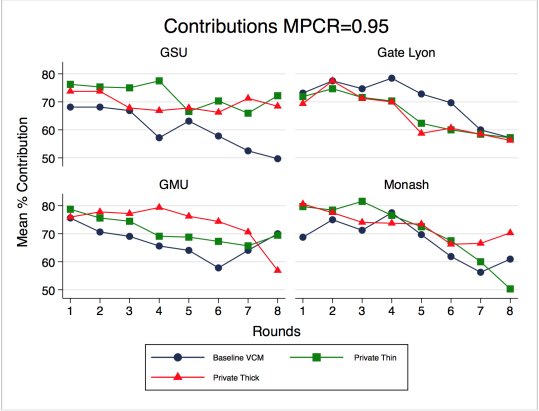



*Table 1: Average Contributions as Percentage of Endowment, by Treatment and Location*

| | 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, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

*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=0.50* | | | | |
|---|---|---|---|---|
| | Successful | Failed | | |
| | Original study | Replication=1 | Replication=2 | Replication=3 |
| $\pi$ | PSP | | | |
| 0.01 | 0.09 | 0.05 | 0.02 | 0.01 |
| 0.05 | 0.34 | 0.20 | 0.11 | 0.06 |
| 0.1 | **0.53** | 0.36 | 0.22 | 0.12 |
| 0.15 | **0.64** | 0.47 | 0.31 | 0.18 |
| 0.2 | **0.71** | **0.55** | 0.38 | 0.23 |
| 0.25 | **0.77** | **0.63** | 0.46 | 0.30 |
| 0.3 | **0.81** | **0.68** | **0.52** | 0.35 |
| 0.35 | **0.84** | **0.72** | **0.57** | 0.40 |
| 0.4 | **0.87** | **0.72** | **0.57** | 0.40 |
| 0.45 | **0.89** | **0.80** | **0.67** | **0.50** |
| 0.5 | **0.91** | **0.83** | **0.72** | **0.56** |

*Notes*: Table 5 reports the PSP for different priors $\pi$ after one statistically significant original study, and three subsequent failed replications. 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 Comparative Statics Analysis

In this section, we characterize simple comparative statics about the choice of original authors to replicate their study by varying three core parameters of the model: scholars' priors $\pi_R$ about the likelihood that their results replicate; the value scholars place on science, $\tau$; and the level of seniority of scholars, $S$.

We first look at priors $\pi_R$. Consider first a case in which the original authors had very little confidence in the replicability of their own work. In the limiting case where $\pi_R \simeq 0$, the authors would believe that their results would never successfully replicate. In this case, the authors will choose to replicate if and only if the expected value from a collection of failed replications is higher than the expected value from publishing the paper alone. That is, when:

$$S \cdot \delta \cdot [J \cdot (\tau + R + 1) \cdot \phi E] \geq (1 - E) \cdot J \cdot S + p \cdot (-\tau \cdot \delta \cdot C_\tau) + (1 - p) \cdot [(1 - q) \cdot \delta \cdot P \cdot (\tau - R)] \quad (12)$$

By rearranging, we obtain:

$$E_{\pi_R=0} > \frac{1}{\phi \cdot \delta(\tau + R + 1) + 1} + \frac{p(-\tau \cdot C_\tau) + (1 - p)\delta \cdot P \cdot (\tau - R)}{J \cdot S \cdot [\phi \cdot \delta(\tau + R + 1) + 1]} \quad (13)$$

In the opposite limiting case, that is when $\pi_R \simeq 1$, then the authors will replicate using BL if:

$$S \cdot \delta \cdot [J \cdot (\tau + R + 1) \cdot E] \geq (1 - E) \cdot J \cdot S + p \cdot (-\tau \cdot \delta \cdot C_\tau) + (1 - p) \cdot q \cdot \delta \cdot P \cdot (\tau + R)] \quad (14)$$

After rearranging, we obtain:

$$E_{\pi_R=1} > \frac{1}{\delta(\tau + R + 1) + 1} + \frac{p(-\tau \cdot C_\tau) + (1 - p)\delta \cdot P \cdot (\tau + R)}{J \cdot S \cdot [\delta(\tau + R + 1) + 1]} \quad (15)$$

We can notice that all else equal, $\forall \phi < 1, E_{\pi_R=0} > E_{\pi_R=1}$, that is, the lower bound of editors' tastes for replications that makes replications appealing to scholars is higher for scholars with low priors $\pi_R = 0$ compared to scholars with high priors $\pi_R = 1$. This implies that if $E$ is between $E_{\pi_R=0}$ and $E_{\pi_R=1}$, then authors with higher $\pi_R$ will choose to replicate, while authors with lower $\pi_R$ will choose not to replicate.

We now consider variations in the value $\tau$ authors place on producing and disseminating scientifically valid results. In the most general case, authors will choose to replicate if:

$$[S \cdot \delta \cdot [J \cdot (\tau + R + 1) \cdot E] \cdot [\pi_R + (1 - \pi_R)\phi] \geq (1 - E) \cdot J \cdot S + p \cdot (-\tau \cdot \delta \cdot C_\tau) + \quad (16)$$

$$+(1-p)\cdot\{q\cdot[\delta\cdot P\cdot(\tau+R)\cdot\pi_R]+(1-q)\cdot[\delta\cdot P\cdot(\tau-R)\cdot(1-\pi_R)]\}$$

After rearranging this equation, we obtain:

$$E_{\tau>0}\geq\frac{1}{\delta\cdot(\tau+R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1}+ \tag{17}$$
$$+\frac{p\cdot(-\tau\cdot\delta\cdot C_\tau)+(1-p)\cdot[q\cdot(\delta\cdot P\cdot(\tau+R)\cdot\pi_R)+(1-q)\cdot(\delta\cdot P\cdot(\tau-R)\cdot(1-\pi_R))]}{J\cdot S\cdot\{\delta\cdot(\tau+R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1]\}}$$

When authors place no value on scientifically valid results (*e.g.*, $\tau=0$), then this equation reduces to:

$$E_{\tau=0}\geq\frac{1}{\delta\cdot(R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1}+\frac{(1-p)\cdot\pi_R\cdot[q\cdot(\delta\cdot P\cdot R)+(1-q)(1-\pi_R)\cdot(-\delta\cdot P\cdot R)]}{J\cdot S\cdot\{\delta\cdot(R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1]\}} \tag{18}$$

We notice that $E_{\tau=0}>E_{\tau>0}$, that is, the lower bound of editors' preferences making replications appealing is lower for scholars who place little value on science relative to scholars who place a higher value.

Finally, we consider variations in the level of seniority $S$. When $S>0$, the lower bound of editors' value $E$ making replications appealing can be rewritten as:

$$E_{S>0}\geq\frac{1}{\delta\cdot(\tau+R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1}+ \tag{19}$$
$$+\frac{p\cdot(-\tau\cdot\delta\cdot C_\tau)+(1-p)\cdot[q\cdot(\delta\cdot P\cdot(\tau+R)\cdot\pi_R)+(1-q)\cdot(\delta\cdot P\cdot(\tau-R)\cdot(1-\pi_R))]}{J\cdot S\cdot\{\delta\cdot(\tau+R+1)\cdot[\pi_R+(1-\pi_R)\phi]+1]\}}$$

Increasing seniority has an ambiguous effect on the lower bound $E_{S>0}$. Suppose first that the original paper will never be replicated, that is, $p=1$. If scholars place no value on science ($\tau=0$), then increases in seniority will have no effect on the likelihood of replicating. If instead $\tau>0$, then the lower bound of $E$ making replications appealing will be higher for senior scholars than for juniors (*e.g.*, $\frac{\partial E_{S,\tau>0}}{\partial S}>0$). This is because (i) there is no risk of seeing one's paper falsified and (ii) as scholars become more experienced, they also become more capable of publishing without replications on highly ranked journals.

Next, suppose that the original paper will be replicated for sure (*e.g.*, $p=0$). For successfully replicated papers ($q=1$), and for any positive values of science $\tau$ and reputation $R$, increases in seniority will reduce the lower bound $E$ necessary to make replications appealing. This is because scholars become more patient and know that their work will replicate. Differently, for unsuccessfully replicated papers, whether seniority increases or decreases the lower bound $E$ depends on the relative importance scholars place on science $\tau$ and reputation $R$. If scholars value science more than their

own reputation, $\tau > R$, then increases in seniority will reduce the lower bound of $E$, that is, senior scholars will require a smaller lower bound of $E$ to replicate compare to junior scholars. Differently, if scholars value reputation more than science, $\tau < R$, then seniority will increase the lower bound of editors' preferences $E$ necessary to make replications appealing. This is because when seniority increases, the reputation hit is compensated by publications on better ranked journals (remember that in $t = 1$ payoffs are $(1 - E) \cdot J \cdot S$ from publishing without replicating).[38]

## A.2 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

[38]Note that we are implicitly assuming that the reputational hit from failed replications is independent of the quality of the journal. One could alternatively argue that the reputational hit from failed replications of a top journal publication is much larger than the reputation hit from a failed replication of a relatively minor publication.

*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 | 1.5 | 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*: All models report estimates from lineal models with standard errors clustered both at the group and individual level. Model 1 in all samples include only 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. Model 2 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

## Online Appendix B

*Table B1: Average 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 B2: Average 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 B3: Average 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 B4: 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*: All models report estimates from lineal models with standard errors clustered both at the group and individual level. Model 1 in all samples include only 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. Model 2 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 B5: 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 received | -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 received | 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.