To Replicate or Not To Replicate? Exploring Reproducibility in Economics through the Lens of a Model and a Pilot Study

By ZACHARIAS MANIADIS, FABIO TUFANO AND JOHN A. LIST[*]

The sciences are in an era of an alleged 'credibility crisis.' In this study, we discuss the reproducibility of received empirical results, focusing on economics research. By combining theory and empirical evidence, we discuss the import of replication studies, and whether they improve our confidence in novel findings. The theory sheds light on the importance of replications, even when replications themselves are subject to research biases. We then report data from a pilot meta-study of replication in the subfield of experimental economics, which serves as a positive benchmark for investigating the credibility of economics. Our meta-study highlights certain difficulties when applying meta-research and systematizing the economics literature.

[*]Maniadis: Economics Department, School of Social Sciences, University of Southampton, Southampton SO17 1BJ, UK (email: Z.Maniadis@soton.ac.uk). Tufano: CeDEx, School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, UK (e-mail: fabio.tufano@nottingham.ac.uk). List: Dept. of Economics, University of Chicago, 1126 E.59th Street, Chicago, IL 60637, USA (e-mail: jlist@uchicago.edu). We are greatly indebted to Colin F. Camerer and Tom Stanley for their detailed comments and suggestions on previous drafts. We also thank Robin Cubitt, Jacob Goeree, Charles Plott, Uri Simonsohn, Roberto Weber and participants at both the MAER-NET 2015 conference and the 2016 NIBS (Network for Integrated Behavioural Sciences) Autumn Event. We are grateful to Justin Holtz and Katherine Auger for excellent research assistance. The authors have no financial or other material interests related to this research to disclose.

Is the model of self-correcting science and cumulative knowledge growth a fitting description of the contemporary world of academic research? An active debate about whether this is the case has recently developed, both among academics (e.g., Ioannidis, 2005; 2008; 2012) and the popular press. Other critics have joined, arguing that there is a 'credibility crisis' in several scientific disciplines, including psychology (Nosek et al., 2012), management (Bettis, 2012), and several branches of the biological and human sciences (e.g., Jennions and Moller, 2002; Ioannidis, 2005).[1] The word 'crisis' refers to a widespread concern that a sizable fraction of published findings are type-I errors, or 'false positives' (i.e., scientific 'discoveries' of statistical relations that are in fact not true). Given the great importance of science and the amount of resources with which society entrusts it, an excess of false positives constitutes a severe problem.

This credibility concern has led to the development of empirical studies that offer a birds-eye view of the literature. Many of these developments have taken place in psychology, with methodological advances (e.g., Simonsohn, 2016) special issues in elite journals (e.g., Pashler and Wagenmakers, 2012) and large scale collaborations (Open Science Collaboration, 2012) having been devoted to the topic. This, in its turn, has led to a heated debate (e.g., Gilbert et al., 2016) regarding the state of the discipline. Given the import of the potential problem and the substantial methodological differences of economics from biomedical disciplines and psychology, one may ask: are these techniques and insights transferable to economics? And, should they be applied more frequently?

Importantly, the relative degree to which economics research suffers from the reproducibility problem is still not fully understood. Camerer et al. (2016) replicated several well-known experiments in economics following the approach of large-scale replication of

[1] The discussion has been so diverse and multidisciplinary that different fields have been using different terms to refer to the crisis, spurning some confusion. Goodman et al. (2016) solidify the existing discussion by proposing a general terminology to refer to key relevant concepts such as different types of reproducibility.

the Open Science Collaboration. Based on a smaller number of replications, Camerer and colleagues found that experimental economics research published in top journals is more reproducible than psychological research. The crisis from other disciplines and the novel focus on reproducibility have also spurred responses, including the new Berkeley Initiative for Transparency in the Social Sciences and the launch of the *Journal of the Economic Science Association* devoted to methodology and replication. In addition to proposing replicability strategies employed in other disciplines, it is important to examine whether it is possible to employ tools from meta-research to identify and explore the critical dimensions of the reproducibility of economic research.

In this paper we follow this approach. We first add structure to the concept of "measuring the credibility of a research" using an extension of the Bayesian framework employed in Ioannidis (2005) and Maniadis et al. (2014) to derive the Post-Study Probability (PSP) that an association is true. Our framework shows that to assess if we should update our priors upon receiving new results, we should be interested in much more than p-values. Indeed, any appropriate updating requires information not only on received p-values, but also critically depends on research priors and statistical power of the experiment. We illustrate how the nature of the inference problem, given current practices, can lead to unreliable results and we discuss the lack of quantification of research priors and statistical power in economics. But, even if initial results are unreliable, does replication not ensure that the economics literature is credible as a whole?

We develop a model of replications that allows us to pinpoint the conditions for replications to succeed in safeguarding the credibility of our discipline. Indeed, replication has broad reach, as it can be powerful even when replication research is itself subject to research biases. Our model shows that, naturally, a research design with large average power lends itself to faster convergence to the truth. Perhaps surprisingly, this entails that in the

presence of a set of replication results containing both successes and failures our posteriors can be particularly low if the relevant designs had high average power.

We then perform a pilot meta-study of experimental economics, the sub-discipline of empirical economics with arguably the best *a priori* credibility (Harrison and List, 2004; Duflo, 2006; Angrist and Pischke, 2010).[2] Our approach is to emulate a meta-research model from psychology (Makel et al., 2012). Our empirical attempt reveals substantial difficulties, stemming from the lack of a tradition in meta-research. Developing such a tradition will enhance the potential of quantitative methods to study reproducibility practices and outcomes in economics.[3]

The remainder of the paper is organized as follows. Section 1 introduces the basic framework of analysis and provides thought examples. Section 2 discusses the existing evidence in economics. Section 3 presents the model of biased replication and its implications for inference. Section 4 describes our empirical study and presents its results. Section 5 addresses what our empirical results entail within the model of replication. Section 6 concludes with a discussion of future research avenues.

1. Newly Discovered Associations: A Methodological Appraisal

Given publication of a newly discovered finding, how much confidence should we have that it is true? Following Maniadis et al. (2014), we use a framework developed in the life sciences (Wacholder et al., 2004; Ioannidis, 2005) to assess the fraction of findings corresponding to false positives; and, therefore, to derive a measure of the confidence with which we should view empirical results. The model pertains to classical experiments with

[2] In our empirical study, we focus on the sub-discipline of experimental economics, but our approach speaks generally to empirical analyses. In fact, experimental economics is a limiting case of empirical economics in which researchers have higher degrees of control on the data-generation process. Therefore, it can constitute the basis for comparison to other empirical fields in economics. For Ioannidis and Doucouliagos (2013) "experimental designs have inherently better protection from many confounding biases than observational data", a view that recent empirical evidence appears to corroborate by reporting that experimental economics exhibits abnormal patterns of statistical results to a smaller degree than other empirical fields in economics (Brodeur et al., 2016).

[3] In this issue Ioannidis, Stanley and Doucouliagos (2016) illustrate the potential of meta-research: they find that the median statistical power in economics is at most 18% and that the typical reported result is exaggerated by 100%.

simple linear hypotheses, which are not only easy to model, but also whose design facilitates replication attempts.

We denote n as the number of research questions[4] examined in a specific field, πn of which are actually true.[5] Using standard terminology, α is the typical significance level in the field (usually $\alpha = 0.05$) and $1 - \beta$ is the typical power of an experimental design in this field. We can think of the process with which 'Nature' determines whether two phenomena are associated with each other as a random experiment. Using this interpretation, π is a probability, and we are interested in the Post-Study Probability (*PSP*) that a research finding is true. Maniadis et al. (2014) show that *PSP* is given by dividing the number of true associations declared true by the number of all associations which are declared true:

$$PSP = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)} \qquad (1)$$

It is natural to ask: what factors can affect the *PSP* and, in general, our confidence in the published literature? Several factors are relevant for a given discipline: (i) research priors and the existence of structured theory testing, (ii) study power and sample size, (iii) biases and conflicts of interest, (iv) researcher competition in the presence of publication bias, and (v) the frequency of replication studies. Point (iv) is discussed in Maniadis et al. (2014) and (v) will be tackled thoroughly in our main model. We shall now expand on the first three factors.

1.1. *Priors and Theory Testing vs. Exploration*

From equation 1, a key determinant of the *PSP* is the level of the research prior, π. To illustrate the importance of this variable, let us start from the hypothetical case where for a

[4] Each question examines an 'association' or 'relationship' between variables, and thus we shall be using the terms 'association', 'relationship' and 'research question' interchangeably.

[5] π can also be defined as the prior probability that the alternative hypothesis H_1 is actually true when performing a statistical test of the null hypothesis H_0 (see Wacholder et al., 2004): that is, $\pi = \Pr\{H_1 \; is \; true\}$.

given scientific field, 'surprise' discoveries (rather than theory-driven ones) are necessary in order to achieve publication in a major academic journal. For simplicity, assume that the research design has maximum power (β = 0) and the usual $\alpha = 0.05$. Further, assume that 'surprise' associations are those findings that are "1 in 100 results." That is, of 100 potentially surprising associations studied, one is known to be true ($\pi = 0.01$). In this case, the probability of both a true association and a rejection of the null hypothesis is 0.01*1 = 0.01. And, the probability that there is both no association and rejection of the null hypothesis is 0.99*0.05 = 0.0495. Thus, the total probability of rejection is 0.01 + 0.0495 = 0.0595. Thus, when a researcher rejects the null hypothesis in this case, there is an 84 percent (0.04995/0.0595) chance that there is no association (a false positive). Accordingly, there is only a 16 percent chance that the statistically significant finding represents a true association.

The question of where exactly these priors come from is a fundamental problem for Bayesian analysis. One possible answer is straightforward: it is past scientific knowledge that informs our priors for future explorations. In this way, theory is the bridge between existing knowledge and predictions about the results of future studies. Accordingly, in disciplines where empirical studies are based on well-grounded theory, we should expect that research priors are "better defined" (i.e., theory driven research has higher priors, ceteris paribus). On the other hand, in disciplines where exploration and 'discovery' tends to play a larger role than theory, we should expect more sceptical priors.

Of course, for a given research question, theory cannot change the probability that the association is indeed true, so some explanation is in order. An example can illustrate the argument that theory-driven research is associated with more informed priors. In genetic epidemiology there is a huge number of genes that may potentially be associated with some health conditions. Accordingly, since empirical studies are typically not theory-driven, meta-

researchers have made the claim that studies that uncover very strong statistical associations should be seen with much skepticism and their findings be assigned a very low *PSP* (Ioannidis, 2005). The reason is that, since there is no a priori reason to expect one particular gene plays the specific role discovered by a given study, any reasonable prior for the association should be in the order of one in the millions.

On the other hand, provided there is a plausible biological mechanism that can be used to pinpoint the specific gene that is being examined in a given epidemiological study, the prior should be associated with the plausibility of the theoretical mechanism and not just with the number of possible genes. In general, the existence of a tradition of theory-testing tends to change the set of questions that are being examined in a discipline, causing a selection effect of focusing on more plausible associations. [6]

1.2. *Power and Sample Size*

Equation (1) suggests that a research finding is more likely to be true than false if $(1-\beta)\pi/(1-\pi) > \alpha$: that is, for a given significance level the pre-study odds are beneficial and the study is powerful enough. This conveniently allows us to express *PSP* as a function of sample size via the power of the experimental design. Under several simplifying assumptions, List et al. (2011) show that the optimal sample sizes for an experimental control (N_0^*) and treatment (N_1^*) group are equal to $N_0^* = N_1^* = N^* = 2\left(z_{\alpha/2} + z_{\beta}\right)^2 \left(\frac{\sigma}{\delta}\right)^2$. [7] Here,

[6] The direct channel - through higher priors - is not the only channel through which theory-testing translates into a high *PSP*. We also hypothesize that rigorous theory beneficially interacts with other critical variables that will be discussed later. First, let us consider journal editors' preferences. Editors are less likely to have a strong preference for 'positive results' if some submitted paper has a strong theoretical component, for two reasons. First, the theoretical model might in itself be considered a contribution, and second, both confirming and disconfirming formal theories is interesting in its own right. Accordingly, empirical studies with rigorous theory reduce the risk of the results being kept in the 'file drawer.' This should result in less pronounced 'publication bias' for this type of research, all else equal. If the argument that theory-testing results in weaker preference for positive results (captured by the variable μ in Section 1.3 below) is true, this could also attenuate the link between 'research degrees of freedom' and 'research bias.' The absence of rigorous standards and established practices will be more detrimental whenever authors have much to gain from a particular type of result. Put differently, theory-driven (as opposed to exploratory) research in economics ameliorates a 'conflict of interest' inherent in many of the incentives that academics face.

[7] These optimal sample sizes are the smallest sample sizes that achieve a given power level $(1-\beta)$ for a fixed significance level α and variance σ^2. The researcher makes a two-sided test of the hypothesis that $\mu_0 = \mu_1$, and specifies a minimum scientifically relevant level of the treatment effect ($\delta = \mu_1 - \mu_0 > 0$). The estimation of power is based on the sampling distribution that occurs when the true effect equals δ. In our calculations here we assume that the experimental treatment might affect the mean, but not the variance of the outcomes. Furthermore, the sample size is assumed to be large enough, such that the normal distribution is a good approximation to the t distribution that is typically used in hypothesis testing.

σ^2 is the conditional variance of the outcome of both control and treatment groups, δ is the minimum economically relevant difference between mean control and treatment outcomes, and z_p is the p-quantile of the standard normal distribution.

If one follows the literature and uses a significance level of 0.05, and sets experimental power to 0.80, we have $z_{\alpha/2} = 1.96$ and $z_\beta = 0.84$ from standard normal tables. However, we may also consider lower power levels, say 0.50 or 0.20, which can be observed in the literature (Zhang and Ortmann, 2013; Ioannidis et al., 2017). Then, the z_β from standard normal tables are 0.00 and -0.84. In order to detect a one (resp. one-half) standard deviation change in the outcome variable, this would lead to sample sizes of $N^* = 8$ (resp. $N^* = 31$) with a 0.50 power and $N^* = 3$ (resp. $N^* = 10$) with a 0.20 power. This is in contrast with the sample size $N^* = 16$ (resp. $N^* = 63$) that we would need in order to achieve power equal to 0.80. To obtain a clearer idea about the relationship between sample size and the *PSP*, notice that the approximate power function is:

$$1 - \beta = Pr\{Z < -Z_{\alpha/2} \; or \; Z > Z_{\alpha/2} \mid H_1 \; is \; true\} =$$

$$= 1 - \Phi\left(Z_{\alpha/2} - \sqrt{N}\delta/\sqrt{2\sigma^2}\right) + \Phi\left(-Z_{\alpha/2} - \sqrt{N}\delta/\sqrt{2\sigma^2}\right) =$$

$$= \Phi\left(-Z_{\alpha/2} + \sqrt{N}\delta/\sqrt{2\sigma^2}\right) + \Phi\left(-Z_{\alpha/2} - \sqrt{N}\delta/\sqrt{2\sigma^2}\right),$$

with $\Phi(z)$ being the cumulative density function of a standard normal random variable. It follows that *PSP* can be rewritten as:

$$PSP = \frac{\left[\Phi\left(-Z_{\alpha/2} + \frac{\sqrt{N}\delta}{\sqrt{2\sigma^2}}\right) + \Phi\left(-Z_{\alpha/2} - \frac{\sqrt{N}\delta}{\sqrt{2\sigma^2}}\right)\right]\pi}{\left[\Phi\left(-Z_{\alpha/2} + \frac{\sqrt{N}\delta}{\sqrt{2\sigma^2}}\right) + \Phi\left(-Z_{\alpha/2} - \frac{\sqrt{N}\delta}{\sqrt{2\sigma^2}}\right)\right]\pi + \alpha(1-\pi)} \qquad (2)$$

which is increasing in *N*. Thus, as *N* becomes larger, *PSP* increases. That is, from equation 2 we learn that we should be more secure in our initial findings with larger sample sizes, and by specifying the terms of this equation we may gain quantitative insights by just how much.

From this, one might conclude that if a treatment effect is statistically significant with *N* observations, it will necessarily be statistically significant with $N + 1$ observations. The model highlights that this common intuition is flawed. What is true is that as *N* increases we are more likely to find the *truth*—that is because as *N* grows the empirical estimate approaches the truth. The implication is that if a research study rejects a true null (reports a false positive), it is not more likely to reject a true null with a larger sample size. In fact, it is *less* likely to report a false positive because a larger sample size increases the chances of finding the truth.[8]

As we shall show later in our model with biased replication (which abstracts away from publication bias), these insights are not generally true when we conduct inference from multiple replications and evidence is mixed. In ex-post inference, if we condition on a set of results that have positive as well as negative replications, the negative ones are extremely unlikely to be obtained under a true association if studies are well-powered.

1.3. *Research Bias*

To all the factors described above, one should add the subjective element that stems from the fact that study design and empirical data analysis involves a series of subjective choices, which are likely to bias research outcomes whenever the interests of researchers favour a specific pattern of results. This does not necessarily entail fraud, but most often is caused by

[8] If one extends the simple thought experiment presented in Section 1.1 to cases of empirical designs that are less than maximally powered, the results become quite stark. For instance, consider the case of a study with a 0.20 power, which, as the evidence indicates, might not be very uncommon. Using the assumptions on the prior above, we have that the probability of both a true association and a rejection of the null hypothesis is 0.01*0.2 = 0.002. And, the probability that there is both no association and rejection of the null hypothesis is 0.99*0.05=0.0495. Thus, the total probability of rejection is 0.002 + 0.0495 = 0.0515. When a researcher rejects the null hypothesis in this case, there is a 97 percent (0.04995/0.0515) chance that there is no association. Thus, whereas now there is only a 3 percent chance that the result represents a true association, with a fully powered design there was a 16 percent chance. Again, this comes in contrast with the extended model that will be presented in Section 3, where more power does not always result in higher *PSP* in ex-post inference. The case of a unique successful study is a special case of that model.

a natural human tendency to interpret ambiguous signals in a self-serving way (Babcock and Loewenstein, 1997; Dawson et al., 2002). Ioannidis (2005) defined the 'research bias' u as "the combination of various design, data, analysis, and presentation factors that tend to produce research findings when they should not be produced" (p. 697). Notice that the bias does not refer to chance variability or ideological bias in researchers' beliefs (Maniadis et. al., 2014). In the model, the research bias is captured by assuming that a fraction u of the times where a non-significant outcome should *not* be declared, in fact it *is* declared significant (because of research bias). Hence, the *PSP* in the presence of bias (PSP^{Bias}) will be equal to:

$$PSP^{Bias} = \frac{(1-\beta)\pi + \beta\pi u}{(1-\beta)\pi + \beta\pi u + [\alpha + (1-\alpha)u](1-\pi)} \qquad (3)$$

The derivative of this expression with respect to u is negative if $\pi(1-\pi)[\alpha + \beta - 1] < 0$, which is typically true. The concept of bias is multi-dimensional, and in particular we hypothesise that $u = f(d, \mu, \xi)$. For illustration, in what follows we use the simple functional form $u = d\mu\xi$. The first determinant of the bias, represented by the parameter $0 \leq d \leq 1$, captures the 'degrees of freedom' in research (Ioannidis, 2005). This represents, for instance, whether the methodology in a given discipline is mature and standardized. When $d = 0$, no research bias is possible because some flexibility is necessary for biased research to occur. The second component, represented by the parameter $0 \leq \mu \leq 1$, reflects the degree to which researchers have less opportunities to publish non-significant outcomes, which entails strong preferences for some specific pattern of outcomes (it is not only good methodology that matters). As long as there is no preference for positive results ($\mu = 0$) there can be no distortion of the sort we have assumed here (which simply tends to declare results as significant when they should not be declared as such).

The third component of the bias, represented by the parameter $0 \leq \xi \leq 1$, measures the publication pressure and captures the publish-or-perish culture (or, put differently, represents the returns from publication). The variables μ and ξ mediate the transformation of degrees of freedom into biased research – there is empirical evidence that publication pressures tend to increase the bias (Fanelli, 2010). It is the interaction of these three dimensions that generates the bias; in our illustration, this happens in a multiplicative way: degrees of freedom leave room for biases that materialize if there is both an editorial aversion to papers with non-significant effects and strong incentives for publication.

2. The Existing Empirical Evidence in Economics

Now that we have seen the factors that are likely to affect the credibility of economic results, we examine whether evidence exists in economics that will allow us to characterize these factors. We also consider the issue of whether these factors are quantifiable. In the discussion that follows, we assign a primary role to experimental economics. Angrist and Pischke (2010) prominently advanced the idea that the increasing use of the experimental method protects the credibility of empirical economics. It has been argued that experimental research has prima-facie additional credibility relative to observational studies (Harrison and List, 2004; Duflo, 2006; Angrist and Pischke, 2010; Ioannidis and Doucouliagos, 2013). Furthermore, experiments have an increasingly prominent role in economics: in fact, there has been more than a tenfold growth in the number of experimental studies published in top-five economic journals in the last 40 years (e.g., Nikiforakis and Slonim, 2015). Accordingly, the field of experimental economics represents a positive benchmark for investigating the credibility of economics.

Let us first consider what we know about research priors: does it seem to be the case that 'surprising results' are necessary in order to publish in economics? Economics tends to be theory driven. In fact, young PhD students are immersed in the realm of formal

(normative) theory much more than other disciplines. This is still true despite a reverse trend in recent years, especially in applied microeconomics, where structural analysis has become somewhat less popular. A movement in experimental economics toward structural work should also be noted (see Della Vigna et al., 2012 and the cites therein). This may have beneficial side effects for the reliability of empirical economics, since the benchmark predictions of standard economic theory serve as a protective firewall against excessively exploratory research. Hence, the predominance of formal economic theory should result in much experimental research going beyond exploration to theory testing, a feature that tends to increase priors in our model.

What does empirical evidence tell us about the degree to which economics hinges upon theory? Card et al. (2011) focused on a particular set of empirical studies, field experiments, and estimated that 68% of field experiments are purely exploratory, lacking any explicit theory.[9] The other evidence concerning priors that we are familiar with is the work of DeLong and Lang (1992), who examine major economics journals to find articles in which the central null hypotheses set forth by the authors was not rejected. Under their model, they estimated that *none* of the unrejected nulls is true. Their main interpretation of this evidence is that economics journals tend to publish null results only if the pre-study priors of a true association are particularly high.

What do we know about the power of empirical designs in economics? In economics there are limited domains where power analysis is discussed, and especially presented formally in papers (one of these exceptions is revealed preference tests: see the review in Andreoni and Harbaugh, 2005).[10] Importantly, it is still possible to estimate power

[9] The difficulties in categorizing what a theory-driven study is need to be emphasized. Many disciplines, such as psychology, employ theory but not explicit mathematical models. In addition, ex post theorizing after the results is a major confound that can make accurate measurement very difficult. Card et al. (2011) use a relatively mechanistic definition of a theory-testing study, considering as 'theory-driven' any study that presents at least one line of formal mathematical modelling.

[10] One reason to expect a low level of power in economics, as List et al. (2011) emphasize, is that power analysis is not appealing to economists. The reason is that our usual way of thinking is related to the standard regression model. This model considers the probability of observing the coefficient that we observed, if the null hypothesis is true. Power analysis explores a different question: if the alternative is true, what is the probability of the estimated coefficient lying outside the confidence interval defined when we tested our null hypothesis?

retrospectively by means of meta-analytic estimates (or, in general, assumptions about the actual effect size). Until now, there seems to have been very little relevant research in economics: the only analysis we are familiar with is Zhang and Ortman (2013), who estimate the post-hoc statistical power for the dictator-game experiments included in Engel's (2011) meta-analysis and find median power equal to 0.25. This (scarcity of) evidence should be juxtaposed with psychology and related disciplines such as marketing, which have a much deeper tradition of retrospective power analysis.[11]

Filling the power gap in economics, in an accompanying paper of this Features issue, Ioannidis et al. (2017) uses 159 economic meta-analyses of 6,700 studies and over 64,000 estimates to assess statistical power. In particular, they find that across these literatures, the median of the median power in each literature is, at most, 18%. In addition, the median proportion of adequately powered studies (power ≥ 0.8) in a given research literature is no more than 10.5%.

Other key parameters include the degree of competition among research teams in economics and publication pressure in economics relative to other fields (in our model, ξ). Unfortunately, we are not aware of any formal study addressing these questions. One would need to resort to anecdotal evidence for insights. On the one hand, the functioning of the economics knowledge system is characterized by long-run uncertainty about the intrinsic value of research and by a funnel-shaped publication system (e.g., Oswald, 2007). On the other hand, there has been a significant creation of new prestigious journals by major economic associations. In addition, many of the aforementioned features of the system are

[11]Cohen (1962) started this line of research, analysing the 1960 volume of the *Journal of Abnormal and Social Psychology*, and found a median power equal to 0.18 for small effect sizes, 0.48 for medium effect sizes (the relevant case according to the author) and 0.83 for large effect sizes. Sedlmeier and Gigerenzer (1989) review the evidence about a series of more than ten studies in the seventies and eighties that retrospectively calculated power following Cohen's lead. These studies were not only in psychology but also in other social sciences such as sociology and marketing. Studies from psychology and the cognate disciplines of communication and education found similar results as Cohen (1962). However, studies in the fields of Journalism, Sociology and Marketing found considerably higher average power for medium effect sizes ($0.71 - 0.89$). Sedlmeier and Gigerenzer (1989) also conducted the same exercise for the 1984 issue of the *Journal of Abnormal Psychology* and found that not a single study out of 54 relevant papers calculated the power of their design or discussed their choice of sample size. For the papers of that issue, the authors found median power 0.44 for medium effect sizes (compare with Cohen's 0.46). Very recently, Bakker et al. (2012) estimated median power equal to 0.35, not very different from the results of Zhang and Ortman (2013).

common to other competitive disciplines, such as bio-medical sciences (e.g., Young et al., 2008). Additional anecdotal evidence indicates that economics is not considered among the most difficult disciplines for advancing the academic ladder, especially in the US. Fresh economics PhDs compete for assistant professor jobs, whereas in the biomedical and natural sciences they need to pass through a series of post docs. In any event, we know very little on ξ.

The field where evidence in economics is not lacking concerns pure publication bias. For instance, Doucouliagos and Stanley (2013) examine publication biases across different economic literatures and show how these biases depend on the degree of competition between schools of thought. This evidence (along with the interdisciplinary comparisons by Fanelli, 2010b), point toward a moderately high value of μ for economics, since there exists disproportionately many positive empirical results. But, notice that hard evidence – for example in the form of a study such as Franco et al. (2014), who utilize a unique preregistration pool to derive an estimate of publication bias – has not been provided yet in economics.

Finally, is there evidence for the degree to which economic researchers are biased at the individual level? The survey evidence due to List et al. (2001) indicates that a non-trivial fraction of economists (about 4%) has committed a serious research crime (such as data falsification) at least once. Feld et al. (2012) report evidence consonant with the insights in List et al. (2001), and reveal that more than 30% of total researchers admit using strategies that would bias behaviour towards reporting positive results (selectively report of findings, stopping experimentation when finding desired results, etc.). These results point to large empirical degrees of freedom and the potential for research bias, although evidence for the publication pressure in economics relative to other disciplines is lacking.

3. Can we Trust Replications to Ameliorate the Credibility Problem?

We have already seen that the Bayesian model points to key parameters that affect the credibility of results: p-values, priors, theory-testing, power, researchers' incentives, degrees of freedom, etc. Yet, unfortunately in economics we do not have enough rigorous evidence about these variables. Still, it might well be the case that the protective guardianship of replications ensures the reliability of received results.

Before moving to our framework, we should take care to address what is actually a replication. In our theoretical model, we employ the second notion of replication as defined in Levitt and List (2009): implementing an experiment under the same protocol as the original experiment to check whether similar results can be obtained using different subjects.[12] However, research bias is likely to affect original research from several angles. Since there should be no presumption that replication is impervious to bias, one can wonder if replication can achieve its role when it is afflicted by biases itself. This begs the need to model explicitly how new evidence helps us update our beliefs in the presence of potentially biased replication research.

When no research bias is at work, Moonesinghe et al. (2007) and Maniadis et al. (2014) show that a few replications are enough for our beliefs to converge to the truth. It is important to emphasize that in the background of this basic model there is publication bias.[13] For their calculations it is assumed that there are at least r successful replications out of n trials, meaning that the other $(n-r)$ studies are potentially unpublished, and therefore we do

[12] To add more detail, in the empirical part we use the similar – but extended – definitions of direct and conceptual replications of Schmidt (2009), summarised by Makel et al. (2012, p. 538): "In a direct replication, the new research team essentially seeks to duplicate the sampling and experimental procedures of the original research by following the same 'experimental recipe' provided in the methods section of the original publication […] In a conceptual replication, the original methods are not copied but rather purposefully altered to test the rigor of the underlying hypothesis. Whereas direct replication examines the authenticity of the original data, in conceptual replication, the replicator tests the construct and not the datum." In this regard Schmidt (2009) states (p. 95): "Whereas a direct replication is able to produce *facts*, a conceptual replication may produce *understanding* [Italics in original]".

[13] With the term 'publication bias' we merely refer to the tendency of journals to prefer results of one particular type. This does not include any type of bias from the part of researchers. This model assumes that all biases at the researcher level (such as selectively reporting results) are lumped up within 'research bias.'

not know if they are positive or negative. We want to work in an environment where there is no such implicit publication bias.

To establish this environment, we assume that it is known that there are exactly n replications of each study, and all results from these replications are published. Of course, we are fully in agreement with the concept that publication bias afflicts replication studies, but we focus on inference that can be made in the presence of replication that is biased at the researcher level.

We illustrate our ideas in four models of replication. In the first, our benchmark model, only unbiased replication takes place. In the second, only biased replication exists, with a bias aligned with the original result (we call this an environment of 'sympathetic replications'). According to the third model of 'adversarial replication,' replication attempts are biased towards providing evidence opposing the original finding. The fourth, and final, model assumes that there are some unbiased, some sympathetic, and some adversarial replicating researchers. We are interested in examining how each of these replication models helps us to converge to the truth in terms of our posteriors. Does the existence of biases entail that replication can no longer serve its purpose? How does it depend on the environment?

Of course, as Pfeiffer et al. (2011, p. 1) note: "[t]he patterns of bias can be complex and may also depend on the timing of the research results and their relationship with previously published work." Hence, we will work with simplifying assumptions that will deliver tractable models that provide useful insights. We assume that there is an initial positive result having declared the existence of an association and that it is known that there is a fixed number n of replication attempts. Consider the random variable X, namely the

number of successful replication studies,[14] distributed according to the Binomial distribution with probability of success p and number of trials n. The probability of finding r successful replications out of n trials is given by:

$$b(p,r,n) = Pr\{X = r\} = \binom{n}{r} p^r (1-p)^{n-r}.$$

The key question is: what is the Post-Study Probability (PSP^{rep}) of a given positive finding to be true after r successful replications out of n total replication studies? We are particularly interested in the updating of our beliefs due to replication, relative to the beliefs held after the initial study was published – which are captured by the priors π in our model.

3.1. *Unbiased, Sympathetic and Adversarial Replications*

Conditional on a true association (and given power $1-\beta$), the probability of observing r successes in n replications is $Pr\{X = r\} = b(1-\beta, r, n) = \binom{n}{r} (1-\beta)^r \beta^{n-r}$. Conditional on a false association (and given significance level α), this probability is $Pr\{X = r\} = b(\alpha, r, n) = \binom{n}{r} (\alpha)^r (1-\alpha)^{n-r}$. Accordingly, our posterior after all the evidence is:

$$PSP^{rep} = \frac{[Probability\ of the\ association\ being\ true\ and\ having\ r\ successes\ in\ n\ trials]}{[Probability\ of\ having\ r\ successes\ in\ n\ trials]}$$

$$= \frac{b(1-\beta,r,n)\pi}{b(1-\beta,r,n)\pi + b(\alpha,r,n)(1-\pi)} \quad (4)$$

Recall that π is the probability of the association being true after the original (positive) result has been published.

Now, let us consider the possibility of researcher bias in the conduct of replications. First, a bias in favour of the original result may exist if the replication is made by research teams friendly to the original research team. In accordance with Ioannidis (2005), we assume that in each occasion where the replication would have been declared unsuccessful, it will be

[14] There are many possible criteria that can be used to define a 'successful replication'. For instance, a replication study may be considered successful whenever the statistical method used in the original paper delivers a p-value lower than 0.05 when applied to the replication data (e.g., Camerer et al., 2016).

declared successful (meaning, positive), a fraction $0 < v \leq 1$ of the time, due to the bias. Now, what is the probability of the association being true and declared true in r out of n replications? Conditional on the association being true, it will be declared true in a given replication study either because the study will not make a type-II error, or, if the study falls prey to type-II error, because of the bias v. Hence, the Bernoulli probability of each 'success' must now be given by the 'sympathetic bias' formula: $(1 - \beta) + \beta v$.

Similarly, if the association is false, it will be declared true in a given replication study either because the study makes type-I error, or, if the study avoids type-I error, because of the bias v. Hence, the Bernoulli probability of each 'success' must now be given by $\alpha + (1 - \alpha)v$. In other words, conditional on a true association there is a chance $b[(1 - \beta) + \beta v, r, n]$ of declaring r successful replications out of n total replications and conditional on false there is a chance $b[\alpha + (1 - \alpha)v, r, n]$ of declaring r successes out of n replications. Accordingly, the PSP in the presence of sympathetic replication bias v becomes:

$$PSP_S^{rep} = \frac{b[(1 - \beta) + \beta v, r, n]\pi}{b[(1 - \beta) + \beta v, r, n]\pi \; + \; b[\alpha + (1 - \alpha)v, r, n](1 - \pi)} \quad (5)$$

Now, let us assume that there is an 'adversarial bias' $0 < \omega \leq 1$ in the opposite direction, namely in the direction of *declaring against* the initial result (remember this result was positive). This means that out of all cases where the result would have been declared positive, it is actually declared negative (in other words, declared a failed replication) a fraction ω of the time. What is the probability of the association being true and declared true in r out of n replications? Conditional on the association being true, it will be declared true in a given replication study only if the study does not fall prey to type-II error, and at the same time it avoids the 'adversarial bias'. Hence the Bernoulli probability of each 'success' must now be given by $(1 - \beta)(1 - \omega)$. Similarly, if the association is false, it will be declared true in each replication study only if the study makes type-I error and

simultaneously avoids the 'adversarial bias'. Hence, the Bernoulli probability of each 'success' must now be given by $\alpha(1-\omega)$.

In other words, conditional on the association being true, there is a chance $b[(1-\beta)(1-\omega), r, n]$ of declaring r successful replications out of n total replications, and conditional on a false association there is a chance $b[\alpha(1-\omega), r, n]$ of declaring r successful replications out of n total replications. Accordingly, the PSP in the case of adversarial bias becomes:

$$PSP_A^{rep} = \frac{b[(1-\beta)(1-\omega), r, n] \cdot \pi}{b[(1-\beta)(1-\omega), r, n] \cdot \pi + b[\alpha(1-\omega), r, n] \cdot (1-\pi)} \quad (6)$$

3.2. *Heterogeneity across Replicating Teams*

Now, let us assume that there is heterogeneity across replicating teams. In particular, a fraction $0 < \varphi < 1$ is sympathetic (with bias v as above), a fraction $0 < \psi < 1$ is adversarial (with bias ω as above), and the remaining fraction $(1-\varphi-\psi)$ is neutral. What should the academic community infer on the basis of observing r successful replications out of n total replications?

Let us assume first that the association is true. Then, the probability of a given replication being successful (declaring a positive result) is the weighted sum of the probabilities of positive results for each type of researcher. Accordingly, each replication is a Bernoulli experiment with probability of success $\chi_1 \equiv \varphi \cdot [(1-\beta) + \beta v] + \psi \cdot [(1-\beta)(1-\omega)] + (1-\varphi-\psi) \cdot (1-\beta)$.[15] Now assume the association is false. Then, the probability of a given replication being successful (meaning, declaring a positive result) is the weighted sum of the probabilities of positive results for each type of researcher. Accordingly, each

[15] Of course, by simply interpreting φ and ψ as the probabilities of a given replication having taken place by sympathetic and adversarial authors, respectively, we need not invoke any assumption that replicators are randomly drawn.

replication is a Bernoulli experiment with probability of success $\chi_2 \equiv \varphi \cdot [\alpha + (1-\alpha)v] + \psi \cdot [\alpha(1-\omega)] + (1-\varphi-\psi) \cdot \alpha$.

Therefore, conditional on a true association there is a chance $b(\chi_1, r, n)$ of declaring r successful replications out of n total replications, and conditional on a false association there is a chance $b(\chi_2, r, n)$ of declaring r successful replications out of n total replications. Finally, the PSP in the heterogeneous case becomes:

$$PSP_H^{rep} = \frac{b(\chi_1, r, n) \cdot \pi}{b(\chi_1, r, n) \cdot \pi + b(\chi_2, r, n) \cdot (1-\pi)} \quad (7)$$

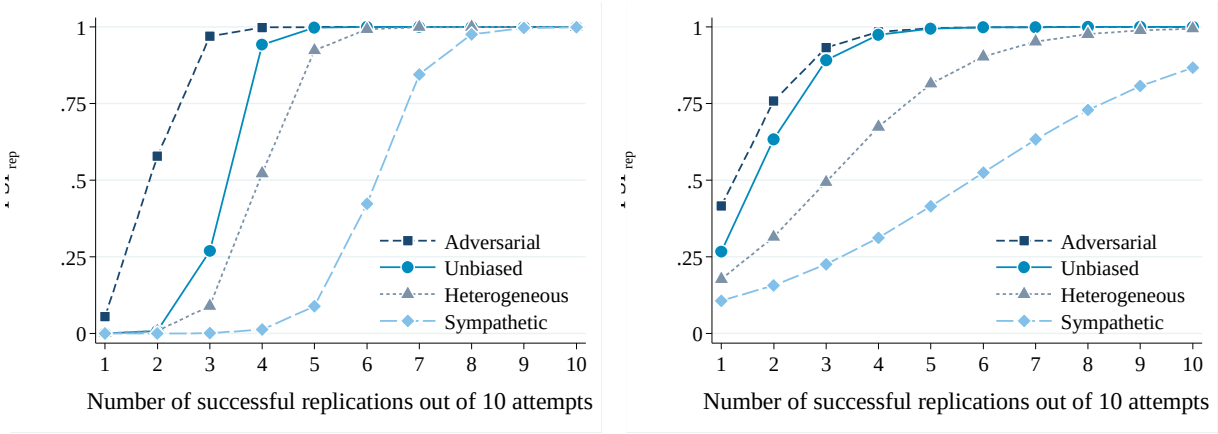
3.3 *Updating on the Basis of Replications*

Figure 1 illustrates the Post-Study Probability as a function of the number of successful replications, when the total number of replications is fixed to 10. That is, we follow our model and plot the *PSP* (denoted PSP^{rep}) for a number of successful replications out of 10 in Figure 1, assuming two different power levels, and $\alpha = 0.05,\ n = 10,\ v = 0.3, \omega = 0.4,$ $\pi = 0.5, \varphi = 0.33,\ \psi = 0.33$. We believe that the specification of Figure 1 is realistic with respect to the prior π, because as Ioannidis (2005) and Maniadis et al. (2014) have shown, it is not unreasonable to expect that after the initial study our beliefs for the truthfulness of the association remain small.

Insights from Figure 1 teach a few useful lessons. First, by comparing the left and right panels, we see the importance of power: regardless of the chosen replication model, power importantly influences the *PSP*. We return to this point below. Second, in disciplines where there is a strong element of competition we might generally expect replications to be adversarial. Hence, our updating on the basis of a fixed number of replications should be larger, relative to a discipline with neutral researchers, and much larger than in disciplines with sympathetic replications.

Accordingly, in disciplines where there is an aversion to contradicting the authors of initial studies we should feel less secure in using the standard Bayesian model when updating – the role of replications for establishing convergence of beliefs to the truth is less powerful. Using similar logic, this means that in a 'sympathetic replicators' environment more successful replications are needed in order to establish the result. For instance, if we wish to achieve PSP of at least 80%, the left panel of Figure 1 shows that seven successful replications are required, whereas only three replications are enough in the 'adversarial replicators' regime. Figure 1 also importantly highlights the need to establish the regime in which replication takes place, since different replication regimes entail very different inference from a given pattern of replication results.

Figure 1. *PSP* as a Function of Number of Replications out of 10 Attempts



a. $(1-\beta)=0.7$ b. $(1-\beta)=0.2$

Note: For the calculations, equations 1 and 5-7 were used while assuming that $\alpha = 0.05,\ n = 10,\ v = 0.3, \omega = 0.4,\ \pi = 0.5, \varphi = 0.33,\ \psi = 0.33$.

Perhaps surprisingly, our model predicts that (for a large range of parameters) conditional on a set of mixed replication results, disciplines where studies are known to be more powerful can be associated with lower *PSP*. To illustrate, let us consider the derivatives of the PSP in each regime. For the unbiased regime the PSP is decreasing in power $(1-\beta)$ as long as $\beta < \frac{n-r}{n}$, for the sympathetic replication regime the same holds if

$\beta < \frac{n-r}{n(1-v)}$ and for the adversarial regime it holds as long as $\beta < \frac{n(1-\omega)-r}{n(1-\omega)}$. This non-monotonicity implies that in an environment of low power (economics is one such environment, according to the results of Ioannidis et al., 2017) low success rates of replications may often be less condemning (in terms of the *PSP*) relative to a high-power environment.

The intuition for this paradox is that more power is always beneficial ex ante, in order to increase the chances of obtaining a positive result. However, at the ex-post stage it is often the case that a given set of replication results is more likely to have originated from a true association if the power is low, rather than high. The reason is simple: in our model we condition on a number of replications that include successes as well as failures. If power is high, numerous failures are very unlikely if the association is true, because the probability of type-II error is low. If the evidence is mixed, it is possible that disciplines with low average power feel more comfortable for the truthfulness of their results. Illustrating this point, the right panel of Figure 1 shows that when (1-β) = 0.2 the PSP for low levels of replication is higher than in the high-power environment (left panel) of Figure 1. In particular, in the 'unbiased' and 'adversarial' regimes, three replications are enough to establish a PSP of at least 0.8.

Anecdotal evidence indicates that despite the importance of replications, incentives to conduct them in economics are low, since they are typically not valued as highly as original contributions. In addition, a replication might not be viewed positively from the point of view of the authors of the initial study. Given these presumed incentives, it is worth exploring whether the policies of top economic journals regarding the availability of data and complementary material are enough to encourage replication studies. It is even possible that actual replications may not be declared as such, especially those that contradict original

findings. These issues require systematic study and we shall now make a first step in that direction, presenting a pilot study that quantitatively assesses replication.[16]

4. A Pilot Study of Replication in Experimental Economics

We have seen how in the absence of relevant evidence about the credibility of initial findings, the role of replication becomes critical. Replication is the cornerstone of the scientific method, and we have shown that it should be considered a safeguard against the false positives problem even if it is practiced with biases, and even if our studies have low power. We now examine empirically the degree to which replication is playing its role for the subfield of experimental economics. We use a large sample of papers from the most prestigious economics journals to provide an overview of the frequency with which replication has been conducted recently in experimental economics. This will help to inform us about the following: on the basis of how many replications do we update our beliefs about initial findings?

As emphasized earlier, we focus on studies of causal analyses in economics with relatively simple hypotheses, typically based on reduced forms (standard experimental techniques). This provides a benchmark to examine replicability, since many other fields in economics employ complex structural analyses which make them difficult to replicate using different data sets, especially since field evidence is constrained by the data availability.

From experiments, experts would argue that they understand the robustness of phenomena, such as preference reversal over a pair of alternatives conditional on the elicitation method, giving in ultimatum games, the decline over time of cooperation in public good games, and the convergence to the equilibrium in competitive markets with classic

[16] Maniadis et al. (2015) discuss possible dis-incentives of initial researchers to have their work replicated, as well as recently proposed solutions. Prominent among those proposals are journals devoted to replications, special grants for funding them, large-scale preregistered replications (e.g. Camerer et al. 2016), (quasi-)adversarial collaborations (e.g., Alempaki et al., 2016) and editors enforcing direct replications as a prerequisite for publication. Maniadis et al. (2015) also propose that incentivizing having one's work replicated could address important shortcomings of the current system.

double auctions design, to mention a few examples. What does a quantitative meta-research study have to offer? First, often expert opinion might not be enough, as evidenced by the fact that in Camerer et al. (2016)'s study, the predictive accuracy of experts about replicability did not outperform objective indices such as sample size and p-value. Moreover, Della Vigna and Pope (2016) find that several dimensions of expertise such as academic rank, citations and even local expertise do not improve predictive forecasting performance for a series of experimental treatments (on the other hand, confidence in one's forecast is moderately correlated with performance). Second, systematic meta-research studies allow the examination of trends in the overall literature and facilitate quantification of critical variables. This approach is complementary to expert opinion and can yield new insights, especially in the current era characterised by enormous, fast-growing, scientific production and increasing specialization. For instance, our pilot study finds that very few experiments in the discipline are declared as replications, a practice that may make it difficult for obtaining unambiguous meta-analytic summaries for certain effects.

There are several additional research questions that we aim to address. What fraction of experimental economic papers in each journal, and in the whole literature, are replications? How does experimental economics fare relative to other fields, such as experimental psychology? [17] It has also been argued that many implicit replications take place in experimental economics, in the form of 'benchmark studies' that emulate prior designs and serve as controls to compare new treatments. To what extent is the existence of these implicit replications true, and how do reporting practices in the discipline affect the meta-researcher's capacity to measure replicability? In addition, how does the fraction and number of replications move over time in the experimental economics literature? Are field

[17] In psychology, Makel et al. (2012) examined the papers published in the top 100 journals (according to the 5-year impact factor) and estimated the percentage of them that represent replications. They found that about 1% are replications, most of which are successful; moreover about 80% of the total replications are 'conceptual' rather than 'exact'. Similarly, less than 2.5% of research articles in marketing are replications or "extensions with replication" (Evanschitzky et al., 2007; Hubbard and Armstrong, 1994). However, when replications are published, they tend to be strong contradictions of the initial findings about 50 percent of the time.

experiments in economics more difficult to replicate relative to lab experiments, due to their cost and complexity? Do other parameters such as authorship overlap, etc., seem to affect the rate of success of replications?

We used EBSCOhost to search into EconLit database (searching in EconLit via EBSCOhost appears more easily manageable relatively to accessing EconLit directly). We restricted our searches to the top 150 journals in economics according to the *Eigenfactor Score*[18] of '*ISI Web of Knowledge Journal Citation Reports* 2013 Social Science Edition (JCR for short).[19] We first retrieved all documents published in English in the years 1975-2014 in the identified top 150 journals in economics. Looking for experimental studies, we restricted our search among the papers that used the root "experiment*." anywhere in the article. We then focused our attention on the subset of papers that used both the root "experiment*" and "replicat*" in order to analyse whether they represented actual replications of experimental studies. Because the use of the root "replicat*" is neither necessary nor sufficient for a paper to contain a replication, we searched for replications in the totality of experimental papers.[20]

To this end, we also considered papers that used the root "experiment*" but not the root "replicat*". First, we wanted to learn how many of those papers reported actual experiments. Hence, we first randomly selected a stratified[21] sample of 2001 of these papers and examined whether they constituted actual experimental studies. In the next stage of our

[18] The definition from the JCR interface states: "The *Eigenfactor* Score calculation is based on the number of times articles from the journal published in the past five years have been cited in the JCR year, but it also considers which journals have contributed these citations so that highly cited journals will influence the network more than lesser cited journals. References from one article in a journal to another article from the same journal are removed, so that *Eigenfactor* Scores are not influenced by journal self-citation."

[19] A complete list of the top 150 journals can be obtained from JCR 2013 Social Science Edition (as described in the Online Appendix). According to Eigenfactor Score in the 'JCR 2013 Social Science Edition' the *American Economic Review* is ranked 1st, while the *National Tax Journal* is ranked 150th.

[20] As noted above, many papers contain only one 'arm' of a given treatment that appeared in an earlier study. These studies were classified as '*quasi replications'*. Such studies do not add to our knowledge in the same manner as basic replications since there is no actual 'treatment effect' to compare with the original study. We choose to measure the scope of these quasi-replications because they are often invoked as salutary practices, but their existence may give a potentially illusory sense of reproducibility in the discipline.

[21] The total population of papers considered was stratified according to the year of publication. Then, sampling fractions were obtained from the strata such that the sampling fractions would have been proportional to the total population of papers considered.

analysis, in order to verify the existence of "implicit" replication studies,[22] we carefully went through 500 randomly selected studies among those that we categorized as experimental to determine which studies constitute replications.

In the final part of our analysis, we went through all of the papers that we categorized as replications (either having the root "replicat*" or not) to categorize important variables that may affect the degree to which replication studies tend to confirm the original results (we call this 'success rate'). Some of these variables include direct replication (vs. conceptual), number of times the original study has been cited, overlap in the authorship of the original and the replication, the replication being published in the same journal as the original study, and others.[23]

In a recent paper, Duvendack et al. (2015) review all replications in economics and among them are a few experimental studies. However, since their search of replications is not systematic, their evidence does not inform us greatly about the frequency of replications in the economics literature. This is one of the variables that we are quite interested in determining: what is the average number of replications in each empirical study?

4.1. *Results*

In total, 206,522 papers were retrieved, after searching for all papers in English in the top 150 journals according to the *Eigenfactor Score* of JCR. Among them, 8,886 papers contained the term root "experiment", with 1,158 papers containing both the roots "experiment*" and "replicat*." Examining studies using both the roots "experiment*" and "replicat*", we found

[22] Assuming that replication studies may score lower in terms of novelty in the eyes of other scholars, it could be possible that authors may refrain from using the root "replicat*" even when implementing a replication. By contrast, we do not anticipate any reasons that may make authors of an experimental study avoid the use of the root "experiment*"; hence, if authors of experimental studies do not use the root "experiment*" we expect that this would have happened for causes orthogonal to our research questions.

[23] To examine the robustness and consistency of the categorizations of our research assistants, two of the authors conducted - for a sample of articles - the three main tasks: determining which articles contained new experiments (task 1), determining which articles contained a replication (task 2) and coding the individual characteristics of replications (task 3). We drew a sample equal to the greater between 10 papers and 2% of the total number of papers considered for each of the three aforementioned tasks completed by the research assistants. Our consistency checks revealed that some of the categorizations were more demanding than others. In particular, concerning the first task, 95% of results were consistent with the research assistants' categorization. When it comes to task 2, the agreement of authors' judgement with the research assistants is about 81%. Finally, in reference to the third task, agreement between the categorization of the authors and the research assistants was about 70%. Categories that revealed themselves as particularly demanding to agree upon are the categories for 'success' and 'replication type.'

that 654 out of 1,158 papers contained actual experiments. In order to estimate the representation of experiments in the literature as a whole, and to perform our robustness check, we first randomly sampled (in a stratified fashion) 2,001 papers from the 7,754 papers using the root "experiment*", but not "replicat*", and examined them in detail. Only 1,037 of the 2,001 of the sampled papers were found to be actual experiments (see the online appendix for the exact way in which we completed this task).

Second, we estimated the number of experimental papers that are actual replications, among those that contained the root "replicat*" and among those that did not. Among the former 654 papers, 99 turned out to be actual replications. In order to examine the latter, we drew randomly 500 papers among the 1,037 experimental papers that did not contain the root "replicat*." Among those, only 12 were replication papers. Note that in our sample there are 111 replications of which 26 are quasi-replications, which we have excluded from the current analysis.

In the end, we estimated the fraction of total papers in economics that contain new experimental data to be 2.3%.[24] Importantly for our purposes, the fraction of replication studies over the total number of experimental studies is 4.2%.[25] Intriguingly, our estimate is of a similar magnitude as observed in marketing and psychology (Evanschitzky et al., 2007; Makel et al., 2012). An interesting novel aspect of our data relative to Makel et al. (2012) is the fact that in our estimation we also account for 'implicit replication' studies that do not declare themselves as such.

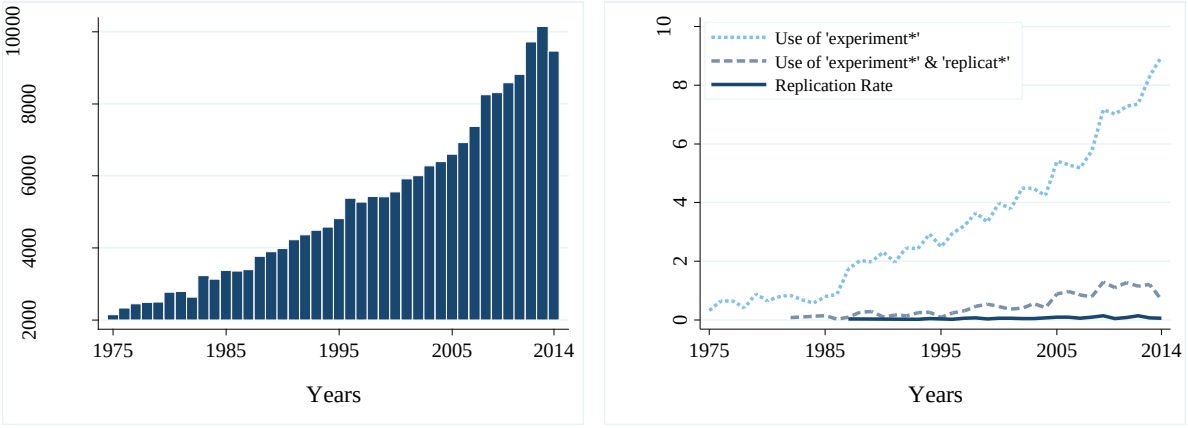
Concerning the replication results themselves, an interesting insight is that we found a 'success rate', based on 85 replications, of 42.3%. This means that roughly 40% of papers that were published as replication studies successfully replicated another experiment. This is

[24] Given that only $1{,}037/2{,}001$ of the sampled papers are actual experiments, we estimate that $4{,}018 = (1{,}037/2{,}001 * 7{,}754)$ of the 7,754 papers using the root "experiment*", but not "replicat*", are actual experiments. Thus, the fraction of total papers in economics that contain new experimental data is estimated to be $2.3\% = [(4{,}018 + 654)/206{,}522] * 100$.

[25] The fraction of *actual* replication papers out of the sample of 500 experimental papers not using the root "replicat*" is $0.024 = (12/500)$. Thus, the fraction of replication studies over the total number of experimental studies (which we estimate equal to $4{,}672 = 4{,}018 + 654$) was estimated to be $4.2\% = [(0.024 * 4{,}018 + 99)/4{,}672] * 100$.

somewhat higher than the outcome of the recent large-scale replication initiative from psychology (Open Science Initiative, 2015) and Duvendack et al. (2015), which both found a success rate of about a third, while, also in psychology, Makel et al. (2012) found a very high success rate of 73%.

Figure 2. Publications and Replication rates in the top 150 journal in Economics according to the *Eigenfactor Score*



a. Number of Publications in top economics journal

b. Use of the roots "experiment*" and "replicat*" along with replication rates in top economics journal

Note: In panel a., the bars represent the total number of articles published yearly in the top 150 economic journals. In panel b., the dotted line reports the percentage of those publications that used the term "experiment*"; the dashed line represents the percentage of those publications that used both the term "experiment*" and the term "replicat*". The solid line describes the total fraction of papers that contain experimental replications relatively to the total yearly publications in the top 150 economic journals.

Figure 2 illustrates the basic trends of economic research over time in our data. The fraction of papers that use the root "experiment*" increases steadily through time (recall that we estimated that about half of those are actual experiments). In addition, the use of the root "replicat*" seems also to increase though time. Notwithstanding, the replication rate appears to be low and fairly stable over time. Table 1 contains detailed information from our set of 85 replications. As can be seen from this descriptive analysis, there is a much larger number of replications in the last 15 years, and the success rate went up slightly. This certainly represents a positive direction for empirical research in terms of replications.

TABLE 1—SUCCESS RATES IN DETAIL

Replication type	Overall	1975-1999	2000-2014
By same authors (N=15)		26.7%	73.3%
Failed	13.3%	0.0%	18.2%
Mixed	40.0%	75.0%	27.3%
Successful	46.7%	25.0%	54.5%
By same journal (N=12)		41.7%	58.3%
Failed	16.7%	0%	28.6%
Mixed	33.3%	60%	14.3%
Successful	50.0%	40%	51.1%
All replications (N=85)		17.7%	82.3%
Failed	11.8%	6.7%	12.9%
Mixed	45.9%	53.3%	43.3%
Successful	42.3%	40.0%	42.8%
Direct (N=45)		11.1%	88.9%
Failed	11.1%	0.0%	12.5%
Mixed	44.4%	80.0%	40.0%
Successful	44.5%	20.0%	47.5%
Conceptual (N=40)		25%	75%
Failed	12.5%	10.0%	13.3%
Mixed	47.5%	40.0%	50.0%
Successful	40.0%	50.0%	36.7%

To delve more deeply into the determinants of replication success we employ Ordered Logit estimations. Our dependent variable is "*Success*," which assumes three values: 2 (replication success), 1 (mixed result), or 0 (replication failure). As explanatory variables, we use the number of citations of the original study (i.e., "*Citation Original Study*") together with dummy variables for identical implementation protocol (i.e., *Same Implementation*), for overlapping research teams (i.e., *Author Overlap*), for the replication having been published in the same journal as the original study (i.e., *Same Journal*), for running the replication in the same country as the original study and for laboratory experiments (i.e., *Laboratory Experiment*).

Results from our estimation are reported in Table 2 as odds ratios.[26] Model 1 indicates that only the variable *Same Implementation* has a significant impact on the dependent variable. In particular, the odds for a successful replication implemented in an

[26] Ordered probit yields very similar estimations and therefore we opted for an ordered logit for its more intuitive interpretation of the estimated coefficients.

identical fashion to the original study are 3.91 times larger than the odds for successful replications when the analyst departs from the protocol used in the original study. Other possible determinants of replication success in our model do not appear to have predictive power at conventional levels.

TABLE 2—ORDERED LOGIT REGRESSIONS OF REPLICATION SUCCESS

Estimation method	Ordered Logit	Backw./Forw. Sel. Ordered Logit
	Odds Ratio	
Dep. variable: Success	*Model 1*	*Model 2*
Citations Original Study	1.001	1.001*
	(0.000)	(0.000)
Same Implementation	3.885**	4.877**
	(2.607)	(3.104)
Author Overlap	0.956	
	(0.579)	
Same Journal	1.023	
	(0.662)	
Same Country	1.543	
	(0.749)	
Laboratory Experiment	1.360	
	(0.800)	
Log-likelihood	-76.890	-77.635
# Obs.	84	84

Notes: *Success* takes value 2 for a successful replication, 1 for a replication with mixed results, and 0 for a failed replication. *Citations Original Studies* counts the number of citations of the original study being replicated (median=106; range=1–4691). *Same Implementation* takes the value 1 if the implementation protocol of the original study and the replication are the same and 0 otherwise. *Author Overlap* has value 1 if the original research team has overlapping members with the replication team, and 0 otherwise. *Same Journal* takes the value 1 if both the original and replication studies are published in the same journal, and 0 otherwise. *Same Country* takes the value 1 if both the original and replication studies were implemented in the same country (0 otherwise). *Laboratory Experiment* reflects the nature of the experiment and assumes value 1 for 'conventional lab,' 0 for 'field' experiments. Standard errors are given in parentheses.

** Significant at the 5 percent level.
* Significant at the 10 percent level.

This comes to a certain extent with surprise given the evidence in Makel et al. (2012), who obtained the intuitive result that having overlapping research teams or similar publishing outlets play a significant role in understanding replication success. Moreover, it is worth noting that even the variable *Laboratory Experiment* does not seem to have any impact on the

chances of success. We find it important to note that the lack of significant effects of our explanatory variables may either be due to a non-existing link between the dependent variable and a given independent variable, but it might also result from the fact that in our econometric analysis there are only 84 observations.[27] Given that the majority of the dependent variables entered in Model 1 do not impact significantly on the chances of success, we implemented both a backward and a forward selection procedure to identify the more parsimonious model.

The two procedures lead to the same model, namely Model 2 in Table 2. This model contains only *Citations Original Study* and *Same Implementation* as explanatory variables. The former variable appears now significant at $p < .10$, but the magnitude of its effect on the odds of success is very small. The coefficient of the variable *Same Implementation* appears to slightly increase in the specification of Model 2, which does not present an important worsening of the log-likelihood.

In summary, the only predictor of replication success that emerges from the econometric analysis of our limited data set is *Same Implementation*. However, despite the fact that citations of the original study do not explain the odds of success of a replication attempt, it may still be interesting to examine the correlations between citations of replication studies, citations of the original studies, and replication success: after all, many view citations as the primary currency for the advancement of scholars' careers. By testing for pairwise correlation, it turns out that citations of replication studies are positively correlated with those of the original study (correlation = 0.3215; p=0.003) and with the success of the replication attempt (correlation = 0.2487; p=0.022).

[27] One observation was dropped in the regression analysis due to missing values for some of the independent variables.

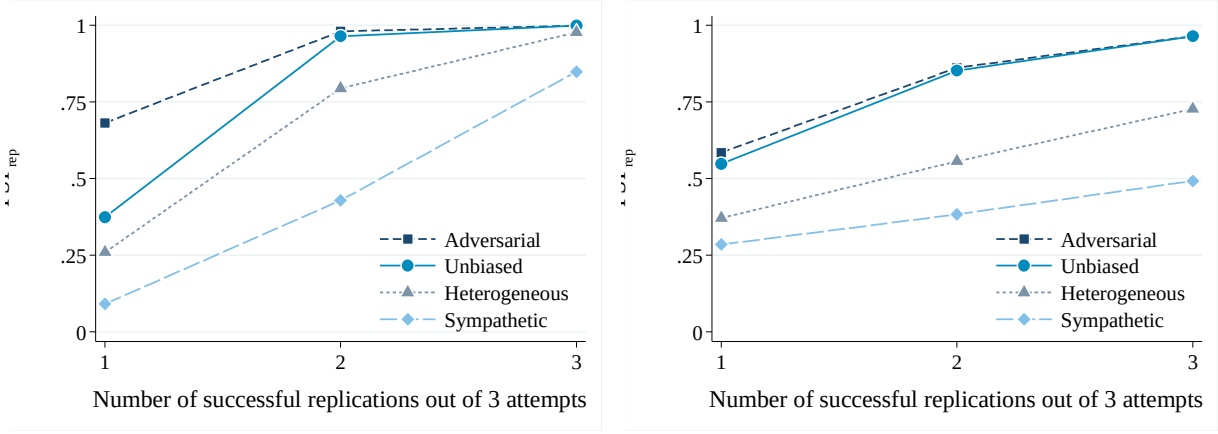
5. A Tentative Meta-Research Assessment of Experimental Economics

What can we say about the experimental economics literature on the basis of our model and empirical results? Unlike clinical trials in medicine, where funding agencies, pharmaceutical companies, and other bodies support independent exact replications, our empirical study reveals that a benchmark of 10 replications is difficult to achieve in experimental economics. In particular, our study reveals that about 4.2 percent of the published experimental papers in economics contain replications. This is somewhat larger than in psychology and marketing (Evanschitzky et al., 2007; Makel et al., 2012), but of similar order of magnitude.

Accordingly, an average of 10 replications in today's research publishing market is too optimistic. Although the number of replications has increased substantially in the last fifteen years, with the incentives for replication today we can think of three total replications as a very optimistic scenario. Figure 3 illustrates this case for the same set of parameters that we examined earlier. As can be seen, even two successful replications out of three are able to establish a very good PSP in our environment, provided that we are not in the 'sympathetic replication' regime. Clearly, additional evidence is needed in order to pinpoint the set of parameters with greater accuracy.

In particular, a key question is whether we can use the empirical evidence to infer what kind of environment we currently reside within experimental economics. This is beyond the scope of our study; however, we argue that survey tools can be devised to elicit the perception of experimental economists concerning the preference of editors for replication studies. In addition, it is possible that by comparing the results of the pre-registered replications with the results of other replication studies we could more fully understand the exact regime in place. Discrepancies may be interpreted as evidence of bias. Since as we have shown the strength of bias depends on the incentives for finding certain types of results, this can inform the possible direction of bias for experimental economists.

Figure 3. *PSP* as a Function of Number of Replications out of 3 Attempts



a. $(1-\beta)=0.7$ b. $(1-\beta)=0.2$

Note: For the calculations, equations 1 and 5-7 were used while assuming that $\alpha = 0.05,\ n = 3,\ v = 0.3, \omega = 0.4,\ \pi = 0.5,\ \varphi = 0.33,\ \psi = 0.33$.

6. Conclusions

In this article, we illustrated the approach of incorporating meta-research tools in examining the credibility of economics research, both theoretically and empirically, by focusing on experimental economics. Our results about study reproducibility suggest that not enough information can be inferred about the credibility of our discipline on the basis of existing evidence, and that replication should not be assumed to automatically work. In the current context of the confidence crisis, economists should develop an awareness of the dangers involved and foster institutions, incentives, and practices that prevent a confidence drift or even strengthen the credibility of economic science. It is now time for economists to understand where we stand, and where we should go in terms of the reliability of our accumulated knowledge.

Increasingly, economists have turned to the experimental model of the physical sciences as a method to understand human behaviour. Much of this research has taken the form of laboratory experiments in which volunteers enter a research lab to make decisions in a controlled environment. Over the past few decades, economists have increasingly made use of field experiments to explore economic phenomena. Whether experimenting in the lab or

field, important open questions revolve around optimal experimentation and inference drawn from the experiments. Accordingly, the degree to which we trust that experimental results are credible is of particular import.

We believe that the work we have done in evaluating the evidence that has accumulated so far does not allow for a general evaluation of the state of affairs in experimental economics. Therefore, instead of proposing solutions to the (possible) crisis in economics, we conclude our paper by emphasizing the need for two types of research: the first concerns the type of research presented in our study. What is the current status of the economics knowledge system? For example, is there bias in the way replication is conducted in economics? In addition, what do we know about the publication system's functioning, incentive structure, prevalent culture, and the diverse stakeholders' influence on the knowledge accumulation? The latter requires research about economists' research biases and conflicts of interest (see Zingales, 2013).

Moreover, other fields have accumulated enough evidence about these issues, and for them the challenge appears to be to correct institutional incentives by designing appropriate rules (Nosek et al., 2012). In fact, proposals for changing the rules of the game have attracted great attention in several scientific disciplines (e.g., Simmons et al., 2011; Landis et al, 2012, Fanelli, 2013, Miguel et al., 2014). Yet, the study of behavioral responses to incentives provided by institutions is beyond the scope of these afflicted disciplines, and therefore the proposed rule changes are not accompanied by rigorous evaluation. This generates a clear and important role for economics to examine the trade-offs and interdependencies of behavior. We therefore call for a more central role of theoretical and empirical mechanism design in attacking this potentially severe problem in modern science.

In this Features issue, Di Tillio et al. (2017) illustrate the power of economic theory by showing how strategic choice of the characteristics of an experimental subject can lead to

biased inference. Importantly, economic theory does not only offer a diagnosis, but it 'dissects the patient' showing what exact types of interventions are likely to work in a given environment. This is fundamentally important before evaluating proposals for reforming science, since as Ioannidis (2012) has forcefully argued: "[…] it is essential that we obtain as much rigorous evidence as possible, including experimental studies, on how these [reform] practices perform in real life and whether they match their theoretical benefits. Otherwise, we run the risk that we may end up with worse scientific credibility than in the current system."

University of Southampton

University of Nottingham

University of Chicago

References

Alempaki, D, Canic, E., Matthews, W., Mullett, T., Stewart, N., Starmer, C. and Tufano, F. (2016). 'Examining how utility and weighting functions get their shapes: A multi-level, quasi-adversarial, replication', Working Paper, University of Nottingham.

Andreoni, J. and Harbaugh, W. (2005). 'Power indices for revealed preference tests', Working Paper, Social Systems Research Institute, University of Wisconsin.

Angrist, J. and Pischke, J.S. (2010). 'The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics', *Journal of Economic Perspectives*, vol. 24(2), pp. 3-30.

Babcock, L. and Loewenstein, G. (1997). 'Explaining bargaining impasse: The role of self-serving biases', *Journal of Economic Perspectives*, vol. 11(1), pp. 109-126.

Bakker, M., van Dijk, A. and Wicherts, J.M.. (2012). 'The Rules of the Game Called Psychological Science', *Perspectives on Psychological Science*, vol. 7(6), pp. 543-554.

Bettis, R.A. (2012). 'The search for asterisks: compromised statistical tests and flawed theories', *Strategic Management Journal*, vol. 33(1), pp. 108-113.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016). 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics*, vol. 8(1), pp. 1-32.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016). 'Evaluating replicability of laboratory experiments in economics', *Science*, vol. 351(6280), pp. 1433-1436.

Card, D., Della Vigna, S. and Malmendier, U. (2011). 'The Role of Theory in Field Experiments', *Journal of Economic Perspectives*, vol. 25(3), pp. 39-62.

Cohen, J. (1962). 'The statistical power of abnormal-social psychological research: A review', *Journal of Abnormal and Social Psychology*, vol. 65(3), pp. 145-153.

Dawson, E., Gilovich, T. and Regan, D.T. (2002). 'Motivated reasoning and performance on the Wason Selection Task', *Personality and Social Psychology Bulletin*, vol. 28(10), pp. 1379-1387.

Della Vigna, S., List, J.A. and Malmendier, U. (2012). 'Testing for Altruism and Social Pressure in Charitable Giving', *Quarterly Journal of Economics*, vol. 127(1), pp. 1-56.

Della Vigna, S. and Pope, D. (2016). 'Predicting Experimental Results: Who Knows What?' Working paper, National Bureau of Economic Research No. w22566.

DeLong, J.B. and Lang, K. (1992). 'Are all economic hypotheses false?', *Journal of Political Economy*, vol. 100(6), pp. 1257-1272.

Di Tillio, A., Ottaviani, M. and Sorensen, P.N. (2017). 'Persuasion Bias in Science: Can Economics Help?' *Economic Journal.*

Doucouliagos, C. and Stanley, T.D. (2013). 'Are all economic facts greatly exaggerated? Theory competition and selectivity', *Journal of Economic Surveys*, vol. 27(2), pp. 316-339.

Duflo, E. (2006). 'Field Experiments in Development Economics', in (R. Blundell, W.K. Newey and T. Persson, eds.), *Advances in Economics and Econometrics Theory and Applications, Ninth World Congress*, vol. 2, pp. 322-348, Cambridge: Cambridge University Press.

Duvendack, M., Palmer-Jones, R.W. and Reed, W.R. (2015). 'Replications in Economics: A Progress Report', *Econ Journal Watch*, vol. 12(2), pp. 164-191.

Evanschitzky, H., Baumgarth, C., Hubbard, R. and Armstrong, J.S. (2007). 'Replication research's disturbing trend', *Journal of Business Research*, vol. 60(4), pp. 411-415.

Fanelli, D. (2010). 'Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data', *PLoS ONE*, vol. 5(4), pp. 1-7.

Fanelli, D. (2010). ‘‘Positive’ results increase down the Hierarchy of the Sciences’, *PLoS ONE*, vol. 5(3), e10068.

Fanelli, D. (2013). ‘Redefine misconduct as distorted reporting’, *Nature*, vol. 494(7436), p. 149.

Feld, L.P., Necker, S. and Frey, B.S. (2012). ‘*Scientific Misbehavior in Economics-Evidence from Europe’,* Working Paper, Walter-Euken Institute and University of Freiburg.

Franco, A., Malhotra, N. and Simonovits, G. (2014) ‘Publication bias in the social sciences: Unlocking the file drawer’, *Science*, vol. 345(6203), pp. 1502-1505.

Gilbert, D.T., King, G., Pettigrew, S. and Wilson, T.D. (2016). ‘Comment on “Estimating the reproducibility of psychological science”’, *Science*, vol. 351(6277), pp. 1037a-1037b.

Goodman, S.N., Fanelli, D. and Ioannidis, J.P.A. (2016). ‘What does research reproducibility mean?’ *Science Translational Medicine*, vol. 8(341), 341ps12, pp. 1-6.

Harrison, G.W., and List, J.A. (2004). ‘Field experiments’, *Journal of Economic Literature*, vol. XLII, pp. 1009-1055.

Hubbard, R. and Armstrong, S.J. (1994). ‘Replications and Extensions in Marketing - Rarely Published but Quite Contrary’, *International Journal of Research in Marketing*, vol. 11(3), pp. 233-248.

Ioannidis, J.P.A. (2005). ‘Why Most Published Research Findings are False’, *PLoS Medicine*, vol. 2(8), pp. 1418-1422.

Ioannidis, J.P.A. (2008). ‘Why Most Discovered True Associations Are Inflated’, *Epidemiology*, vol. 19(5), pp. 0696-0701.

Ioannidis, J.P.A. (2012). ‘Why science is not necessarily self-correcting’, *Perspectives on Psychological Science*, vol. 7(6), pp. 645-654.

Ioannidis, J.P.A., Fanelli, D., Dunne, D. and Goodman, S.N. (2015). ‘Meta-research: evaluation and improvement of research methods and practices’, *PLoS Biol*, vol. 13(10), p. e1002264.

Ioannidis, J. and Doucouliagos, C. (2013). ‘What’s To Know About The Credibility Of Empirical Economics?’, *Journal of Economic Surveys*, vol. 27(5), pp. 997-1004.

Ioannidis, J., Doucoulagos, C. and Stanley, T.D. (2017). ‘The Power of Bias in Economics Research’, *Economic Journal.*

Jennions, M.D. and Moller, A.P. (2001). ‘Relationships Fade with Time: a meta-Analysis of Temporal Trends in Publication in Ecology and Evolution’, *Proceedings of the Royal Society of London*, vol. 269(1486), pp. 43-48.

Landis, S.C., Amara, S.G, Asadulah, K., Austin, C.P, Blumenstein, R., Bradley, E.W.,

Crystal, R.G., Darnell, R.B., Ferrante, R.J. , Fillit, H., Finkelstein, R., Fisher, M. Gendelman, H.E., Golub, Goudreau, J.L., Gross, R.A., Gubitz, A.K., Hesterlee, S.E., Howells, D.W., Huguenard, J., Kelner, K., Koroshetz, W., Krainc, D., Lazic, S.E., and Levine M.S. (2012). 'A call for transparent reporting to optimize the predictive value of preclinical research', *Nature*, vol. 490(7419), pp. 187-191.

Levitt, S.D. and List, J.A. (2009). 'Field Experiments in Economics: the Past, the Present, and the Future', *European Economic Review*, vol. 53(1), pp. 1-18.

List, J.A., Bailey, C., Euzent, P. and Martin, T. (2001). 'Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior', *Economic Inquiry*, vol. 39(1), pp. 162-170.

List, J.A., Sadoff, S. and Wagner, M. (2011). 'So you Want to Run an Experiment, Now what? Some Simple rules of Thumb for Optimal Experimental Design', *Experimental Economics*, vol. 14(4), pp. 439-457.

Loewenstein, G. (1999). 'Experimental Economics from the Vantage-Point of Behavioral Economics', *Economic Journal*, vol. 109(453), pp. 25-34.

Makel, M.C., Plucker, J.A. and Hegarty, B. (2012). 'Replications in Psychology Research How Often Do They Really Occur?', *Perspectives on Psychological Science*, vol. 7(6), pp. 537-542.

Maniadis, Z., Tufano, F. and List, J.A. (2014). 'One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects', *American Economic Review*, vol. 104(1), pp. 277-290.

Maniadis, Z., Tufano, F. and List, J.A. (2015). 'How to Make Experimental Economics Research More Reproducible: Lessons from Other Disciplines and a New Proposal', in (C. Deck, E. Fatas and T. Rosenblat, eds.), *Research in Experimental Economics, Volume 18: Replication in Economic Experiments*, pp. 215-230, Bingley: Emerald Group Publishing.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U. and Van der Laan M. (2014). 'Promoting Transparency in Social Science Research', *Science*, vol. 343(6166), pp. 30-31.

Moonesinghe, R., Khoury, M.J. and Janssens, C.J.W. (2007). 'Most Published Research Findings are False-but a Little Replication goes a Long Way', *PLoS Medicine*, vol. 4(2), pp. 0218-0221.

Nikiforakis, N. and Slonim, R. (2015). 'Editors' Preface: introducing JESA', *Journal of Economic Science Association*, vol. 1, pp. 1-7.

Nosek, B.A., Spies, J.R. and Motyl, M. (2012). 'Scientific Utopia II. Restructuring Incentives and Practices to Promote Truth Over Publishability', *Perspectives on Psychological Science*, vol. 7(6), pp. 615-631.

Open Science Collaboration. (2012). 'An open, large-scale, collaborative effort to estimate the reproducibility of psychological science', *Perspectives on Psychological Science*, vol. 7(6), pp. 657-660.

Oswald, A. (2007). 'An Examination of the Reliability of Prestigious Scholarly Journals: Evidence and Implications for Decision-Makers', *Economica*, vol. 74(293), pp. 21-31.

Pfeiffer, T., Bertram, L. and Ioannidis, J.P.A. (2011). 'Quantifying Selective Reporting and the Proteus Phenomenon for Multiple Datasets with Similar Bias', *PLoS ONE*, vol. 6(3), p. e18362.

Schmidt, S. (2009). 'Shall we really do it again? The powerful concept of replication is neglected in the social sciences', *Review of General Psychology*, vol. 13, pp. 90–100.

Sedlmeier, P. and Gigerenzer, G. (1989). 'Do studies of statistical power have an effect on the power of studies?', *Psychological Bulletin*, vol. 105(2), pp. 309-316.

Simonsohn, U. (2015). 'Small telescopes detectability and the evaluation of replication results', *Psychological Science*, vol. 26, pp. 559-569.

Simmons, J.P., Nelson, L.D. and Simonsohn, U. (2011). 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', *Psychological Science*, vol. 22(11), pp. 1359-1366.

Stanley, T.D. (2001). 'Wheat from Chaff: Meta-Analysis as Quantitative Literature Review', *Journal of Economic Perspectives*, vol. 15(3), pp. 131-150.

Wacholder, S., Chanock, S., Garcia-Closas, M., El-ghormli, L. and Rothman, N. (2004). 'Assessing the Probability that a Positive Report is False: An Approach for Molecular Epidemiology Studies', *Journal of the National Cancer Institute*, vol. 96(6), pp. 434-442.

Young, N., Ioannidis, J. and Al-Ubaydli, O. (2008). 'Why Current Publication Practices may Distort Science', *PLoS Medicine*, vol. 5(10), p. e201.

Zhang, L. and Ortmann, A.. (2013). 'Exploring the Meaning of Significance in Experimental Economics', Australian School of Business Research Paper No. 2013 ECON 32.

Zingales, L. (2013). 'Preventing Economists' Capture', *Chicago Booth Research Paper* 13-81.