The replication crisis, the rise of new research practices and what it means for experimental economics

Written by

Lionel Page

Muriel Niederle

Charles N. Noussair

Robert Slonim

Submitted to the ESA Executive Committee April 2021

In the wake of the replication crisis in psychology, a range of new approaches has been advocated to improve scientific practices and the replicability of published studies in behavioural sciences. The ESA Executive Committee commissioned an ad hoc committee to review the issues raised by the replication crisis, how they affect research in experimental economics, and to make recommendations for experimental economics.

The present report is the result of this review. Its content has greatly benefited from the personal views and insights of a large number of ESA members. The views in the community of researchers in experimental economics are diverse. The present report does not aim at determining a strict ESA policy. Rather, it aims to bring to the community of experimental economists the collective wisdom, which is spread across experimentalists. The report presents an informed discussion of the different issues related to replicability and discusses the different solutions, with their benefits and pitfalls.

The report also contains a series of recommendations which aim to address the challenges presented by the replication crisis, while respecting the diversity of views within the ESA community.

1. The issues raised by the replication crisis

Science is a human practice, and researchers and editors have incentives, which may, in some instances, favour outcomes at odds with the ideal process of scientific discovery. The replication crisis can be seen as a consequence of these incentives and their implications for the use of questionable research and publication practices.

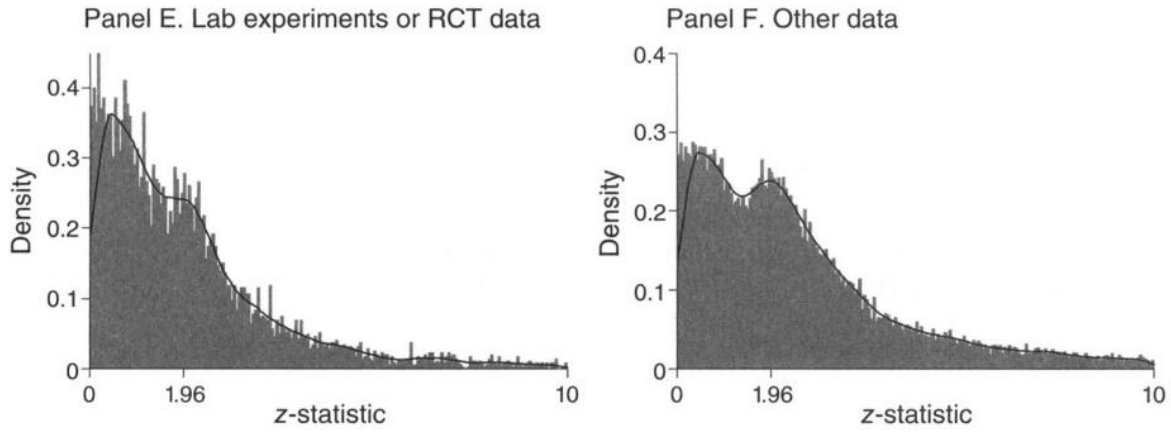
Authors have an incentive to publish in high impact journals; and editors have an incentive to publish findings which are cited more (to raise measures of the impact of their journal). Statistically significant findings are more likely to be cited than null results (Duyx et al. 2017). As a consequence, authors have an incentive to publish significant findings, and editors have an incentive to accept significant findings. *The rewards from publishing statistically significant findings lead to distorting incentives to find something significant to publish.* It leads to biases in the publication of scientific

results which ideally should be about publishing results that answer interesting questions, regardless of whether they yield significant or null results.

File-Drawer effect: Not enough incentives to publish null results. Studies with null results are harder to publish. Authors anticipate this and often do not write up non-significant results. It results in the *drawer effect*: many null results are not published. This has several negative effects. One of them is the wasteful repetition of studies which do not work, because previous failed attempts are not publicly recorded. Another is a misleading body of reported findings in which variables are thought to have stronger relationships than is justified. The drawer effect can also take place within studies. A research team may run several studies linked to each other and only publish the studies/questions which give significant results while leaving unwritten the studies/questions which did not give significant results.

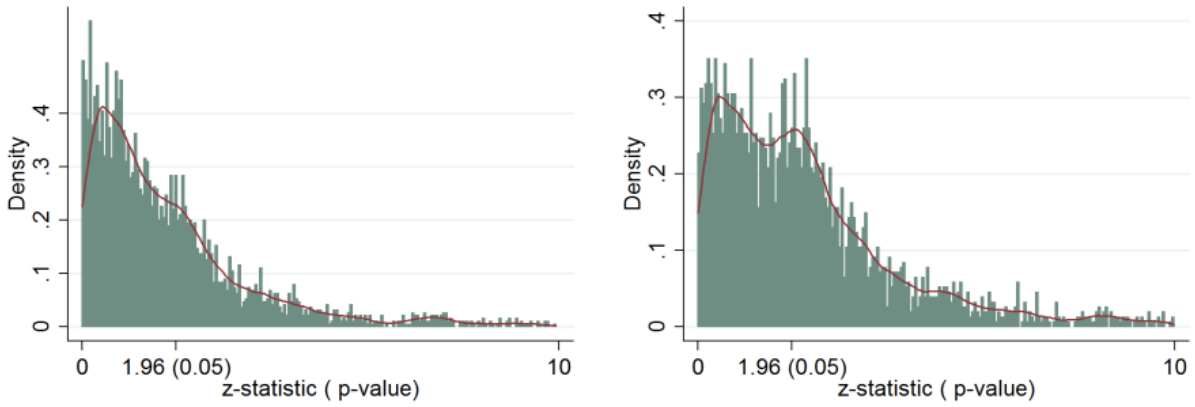
P-hacking: Incentives to mine the data for significant relationships. Researchers have an incentive to use the many degrees of freedom in the available data analyses techniques to find one type of analysis (model, variables, sample), which delivers a significant result. In the case of experiments, a possible form of p-hacking is the systematic investigation of sub-samples. When researchers do not find a significant main effect for their experimental treatment, there is an incentive to look for effects in sub-samples. If researchers were not looking for such effects on sub-samples in the first place, the study's design may not be powered to do so. Besides, researchers may carry tests on many sub-samples and report only the subsamples for which the result is significant.[1]

Brodeur et al. (2016) provide interesting evidence on the prevalence of p-hacking in economics and its sub-disciplines. These figures, taken from the published paper, show that experimental papers display less of a bimodal distribution. Instead, other papers have a clear peak just above z-statistics of 1.96.



The published version lumps laboratory and field experiments together. The working paper version from Brodeur et al. gives a slightly less positive picture for (laboratory) experimental economics than for randomized control trials (RCTs). In spite of the noisier data (likely due to the lower number of observations), there seems to be barely any second peak in RCTs, while the lab experiment distribution seems closer to the non-experimental data. RCTs may be less affected by p-hacking because, given the size and cost of the studies, authors push more to publish even null results.

[1] In such an analysis, they may also fail to correct for multiple hypothesis testing, another widespread problem in the literature.



(a) Randomized control trials data. (b) Laboratory experiments data.

Harking: Incentives to build narratives fitting whatever results are found ex-post. Having found something significant from multiple testing, researchers can reverse engineer "hypotheses" *after the results are known* (HARK). The unexpected results are then presented as an idea which was meant to be studied in the first place.

Compared to other research areas, laboratory experiments have fared relatively well in the wake of the replication crisis (Camerer et al. 2016). There are possibly several reasons for this. One possibility is the existence in economics of a shared theoretical framework, which de facto constrains the type of hypotheses which can be generated to explain results after the fact. Another possible contributing factor is that hypothesis testing is likely more expensive in economics where experiments are incentivised. The possible practice of running many experiments to study many questions, in order to select only a few to focus on ex-post, is less likely to happen in experimental economics than in other disciplines, like psychology, where collecting experimental data is cheaper.[2]

Nonetheless, experimental economics is not immune to the issues raised by the replication crisis. In particular, many experimental papers are increasingly motivated by conjectures which are not driven from economic theory, raising the danger of harking. This is likely a result of the relative decline of the primacy of standard economics theory and the greater openness to ideas coming from other social sciences. Researchers are now able to justify their research with *'reasonable'* assumptions and conjectures reflecting insights from other disciplines such as psychology or sociology where many different opposing hypotheses and views can co-exist. As a consequence, experimental economics is not immune to researchers having many degrees of freedom when designing hypotheses to fit the data.[3] It is especially the case when there is little existing precise theory disciplining the design and a lot of data gathered, such as studies that include extensive psychological measures or physiological data.

2. Pre-registration

The pre-registration of hypotheses has been proposed as one of the tools to improve research practices by reducing researchers' freedom to engage in harking and p-hacking. We discuss separately the pre-registration of hypotheses and the pre-registration of an analysis plan.

[2] This point was made by Noemi Peter in a correspondence.
[3] A point also made by Edward Cartwright.

Pre-registration of hypotheses

Description

The pre-registration of hypotheses requires publishing, on a third party platform, a document describing the intended study and the hypotheses tested. Part or all of the document may be kept private until publication.

Pros

Drawer effect: By providing a record of the study attempted, pre-registration limits the drawer effect. It gives visibility to previous attempts at answering a question.

p-hacking: By specifying the main hypotheses beforehand, pre-registration prevents the ex-post investigation of sub-samples to look for some significant results when the investigated result is not found.

Harking: The pre-registration of hypotheses can limit harking. It puts on record what were the effects expected/investigated. Results which are surprising are more credible when hypotheses are pre-registered. The pre-registration signals that the surprising results that were hypothesized are credible. Effects which are found but were not pre-registered have lower credibility in comparison. Surprising hypotheses which are not pre-registered can be suspected to be ex post rationalisation of surprising results (results which may just be false positives).

Cons

The pre-registration of hypotheses may not deliver the intended benefits.

Pre-registration may reduce the drawer effect but it is unlikely to eliminate it. Even pre-registered, null effects are not published or even written up. They are therefore harder to include in the public record.[4]

Pre-registration creates an additional cost in the process of running and publishing a scientific study. Even if some of the pre-registration work is useful to frontload the conceptual work on an experiment, it still adds some administrative time: writing a document which needs to respect some specific format constraints. It may potentially create some inequality between researchers in large laboratories with junior researchers/research assistants (who often carry the bulk of the administrative work) and researchers in smaller departments (who have the whole responsibility for the design and running of their experiments).[5][6]

The pre-registration of hypotheses also raises the issue of making scientific ideas public before publication. There are two main concerns:

First, ideas could be scooped if the pre-registration is public. One solution is to not require that the pre-registered hypotheses be made public. In that case, it could only be required for the editor and reviewers to have access to it when the paper is submitted to a journal. The pre-registration can be made public when the paper is published.

Second, there could be a risk of "pre-registration trolling" in a way similar to patent trolling. A research group could potentially pre-register a large number of ideas to implicitly lock intellectual

[4] A point made by Irenaeus Wolff.
[5] Points along these lines were made by several people including Noemi Peter.
[6] Egon Tripodi also raised the issue of possible multiple piloting with only successful pilots being pre-registered.

property rights on them without the intent to work on all of them. The payout could be to have the opportunity to select promising projects at a later date as a function of the evolution of the field. One could even envisage that other researchers would offer to co-author these pre-registered studies and carry out the work, in effect paying a rent to the researchers who pre-registered many ideas. While we are not aware of any reports of such practices, it is important to anticipate the types of incentives which could arise from the widespread use of pre-registration. The concern about possible pre-registration trolling raises some questions about the notion that pre-registrations should be seen as establishing property rights. These hazards can be eliminated by making pre-registration private.

Pre-registered Analysis plans

Description

The pre-registration of an analysis plan goes a step further than the pre-registration of hypotheses. It requires the registration of the statistical analysis that is to be used to analyze the experiment data. While often conflated with pre-registration, it is often not required in the pre-registration process. It is, for instance, not required by the AEA RCT registry. A pre-registration with analysis plan can, therefore, be described in short as the pre-registration-plus option.

Pros

Planning. The first merit of the pre-analysis plan may be to force researchers to think more carefully ex-ante about their intended analyses and, therefore, the design. There is potential to improve the quality of experiments being run. In particular, in terms of power which is an important issue, pre-analysis plans may foster better practices with larger samples when needed.

P-hacking. The second merit of the pre-registration of an analysis plan is to limit p-hacking. By clearly specifying the analyses prior to the study, the analysis plan acts as a pre-commitment tool which limits the degrees of freedom of the researchers in the analysis process. As a consequence, it increases the credibility of the results derived from pre-registered analyses. New analyses, not pre-registered, are still possible. They become exploratory and have, by design, a lower degree of strength as evidence.

Cons

Coffman and Niederle (JEP, 2015) pointed out that analyses plans may not dramatically decrease false positives, if there are many competing studies being conducted (either by the same researchers/labs or by different researchers). Intuitively, if 20 analysis plans are written up, then on average one will deliver a significant result under the null. If publications still favour significant results, published results will still overly feature false positives.

There are also risks that pre-analysis plans may constrain the researchers when investigating empirical results. A risk is that “exploratory analyses” end up being too much devalued or dismissed by reviewers and editors. The emergence of such a convention could be detrimental.

In terms of methods, it is dubious to think that there is no insight to gain from looking at the data to determine the best way to analyse it. For a start, researchers often realise that the analyses they initially planned were ill-conceived, or at least not perfectly conceived, after beginning to run the experimental sessions or when looking at the data. One reason for this is that it can be difficult to adequately anticipate all the aspects and challenges which characterise a dataset, before working on it, especially if it is a novel type of experiment.

In terms of results, imposing a strict pre-analysis plan to consider a result as worthy of being considered would prevent researchers from publishing results they did not expect. This does not seem reasonable.

We believe that the view from Simmons et al. (2018) should prevail: "Pre-registration does not restrict the ability to conduct exploratory analyses; it merely allows the researcher and the reader to properly distinguish between analyses that were planned vs. exploratory."[7]

Summary about pre-registration

Given the elements above, the push for a blanket policy of pre-registration of hypotheses and analysis plan does not seem desirable. Letting the academic norms progress with an encouragement for the practice of pre-registration and a monitoring of possible unexpected consequences seems to us the right way forward. The following encouragement can be considered:

Researchers: They could be invited to pre-register their study. The decisions to (1) pre-register and (2) to register hypotheses only or analyses as well can be presented as a choice. Pre-registered hypotheses and analyses would have stronger credibility in the reviewing process, especially in cases such as (but not limited to) theory testing and replication studies. But their credibility may be less well suited to situations where a question/design is novel and more exploratory. Researchers would be free to choose given the trade-offs they face in their specific inquiry. In case of pre-registration, it would be private, but the files should be accessible during the submission process.

Editors, Reviewers: Non pre-registered results should still be considered worthy of publication. However, the pre-registration process could potentially change how much each type of result is valued and how they are discussed, depending on the degree to which the hypotheses are exploratory. All else equal, pre-registered results have stronger credibility. Nonetheless, editors and reviewers should still accept exploratory results, with possibly a more conservative way of appraising them. For instance, greater statistical significance may be important to establish confidence in such results. Alternatively, results may be presented as suggestive and invite further research. We believe that editors and reviewers in experimental economics should refrain from discouraging researchers from publishing exploratory results per se. One possibility would be for editors to propose a standard format for published papers in which tests of hypotheses and exploratory analyses are reported in two different sections of the paper. Another option available to editors, when faced with exploratory results which are quite unexpected, could be to request for some parts of the study to be rerun to ascertain the robustness of the results. This approach cannot be used all the time given the cost of rerunning studies, but in the case of laboratory experiments, rerunning sessions is can be feasible in the revision process.

Registered reports

Description

Registered reports consist of moving the publication point one step back and allowing researchers to submit their pre-registration for consideration for publication. If the registered report is accepted, it seems in practice to work as a strong R&R with high chances of publication for the final study (if it does not deviate too far from the accepted registration).

Pros:

[7] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13*(2), 255–259. View also shared by Anna Dreber and Magnus Johannesson .

Drawer effect. Registered reports have the potential to limit the publication bias on both the editorial and the author side. Editors and reviewers have to decide on the interest of the proposed study before any data is gathered. As editors have a clear incentive to publish significant results, registered reports offer an interesting solution to this issue. On the other hand, editors may still accept papers based on their beliefs about what the results might turn out to be, though this is not an easy task in many instances.

Decision time. An additional benefit of registered reports is that the acceptance decision is frontloaded earlier in the research project. It is better for researchers, in particular, early-career researchers who need evidence of publication success when going on the job market or being considered for promotion or tenure.[8]

In terms of effort, pre-registered reports could in some cases decrease the time spent writing papers. It is not uncommon for authors to re-write papers along different narratives as they progressively investigate their data further. One of these scenarios is when an early working paper with null results becomes a published paper with significant results on a subgroup of the participants or a change in focus to a secondary question. Registered reports force the results to only be written once with a given angle, and this angle is the only aspect of the paper that needs to be addressed with the data once it is collected.[9]

Cons

There are, in theory, possible risks of anti-selection and moral hazard:

Anti-selection: researchers could potentially run several studies and submit as pre-registered reports only those that are not significant. This would contravene the intent of such reports to be registered before any data is collected.

Moral hazard:

1. Researchers could decrease their effort once getting an acceptance so that the quality of the papers published is lower.
2. Researchers could withdraw their papers after acceptance if the results turn out to be very appealing and submit it to a higher-ranked journal. They could even send a paper after acceptance to a higher-ranked journal and then back to the initial journal which offered a registered report if it did not work out at the higher journal.

It seems to us that these risks are not too high given the reputational costs researchers could face. But editors should be aware of them when designing rules for making acceptance decisions for registered reports.

Sealed envelope submissions

An interesting alternative is the idea to submit papers of fully finished studies but where the results are masked. Such so-called "sealed envelope submissions" would induce the editors and reviewers to make their evaluation of the paper's quality and contribution based on the method and on the question rather than based on the results. Martin Dufwenberg and Peter Martinsson (2019), who

[8] We should however point out

[9] This point was made by Irenaeus Wolff. If only a few outlets are offering a registered report publication option, this argument may however not work as authors of rejected submission have then to rewrite their paper to submit it in a traditional format.

proposed this solution, stress that it would reduce the incentives of the researchers to p-hack.[10] This option is appealing by being less costly for the researchers than writing a pre-registered report prior to the study. There is obviously a risk of anti-selection whereby a journal offering this option could receive a disproportionate amount of studies with null results. There is also a question of credibility in preventing information about the results (which already exist) from reaching the editors and referees. Working papers are frequently online, which makes it easy to find the results. Papers are also often presented prior to submission, which may lead to information leakages.

Summary about registered reports

Registered reports are a new and interesting avenue for scientific publications. At the moment, only *Nature Human Behaviour* (NHB) has a registered report section for experimental papers. *Experimental Economics* had a special issue trialling this new type of publication in 2020.

Overall, registered reports appear to be a promising option, and it may make sense to support it for the ESA journals. The use of registered reports would need to be carefully implemented, however. NHB has adopted a registered report option, but it has put the bar much higher in terms of requirements than for its normal paper section, presumably to counter anti-selection. It may, possibly, make it much harder to publish an experimental paper there than through the normal process.

Given that the registration can be either about hypotheses only or about hypotheses and analysis, one could, in the same way, consider two different types of registered report:

Editorial decision based on the registration of hypotheses and design (=pre-submission enquiry+). Researchers propose a design and hypotheses and get an early indication of interest from editors with potential referees.

Editorial decision based on full pre-registration (perhaps including the analysis plan). This is what standard registered reports are.

The editorial commitment for the first type would be weaker than for the second type of pre-registration.

Replications

Description

Replications are the reproduction of published results. We can distinguish exact replication (different teams, same method) and conceptual replications (different teams, different method). We consider here primarily replications of laboratory and online experimental studies, not RCTs in the field which, by their nature (size and context) are harder to replicate.

Pros:

It is widely accepted that replications are highly desirable. They confirm or contradict previous results and helps consolidate the set of accepted results in a field.

Cons

Replications are not free of problems. One possible issue is that the promotion of replications may distort incentives in another way. Replications may be more likely to be published when their results

[10] http://www.u.arizona.edu/~martind1/Papers-Documents/sesebis.pdf

contradict the original studies and are therefore surprising. This publication bias may lead to bias towards researchers overly trying to disprove some past results.

Another problem with replications is that they are often not visible enough. Many so-called zombie results may survive a long time in the collective beliefs of a scientific community while their reality has been undermined by failed replications. This fact also contributes to an asymmetry of incentives to produce good research. Anecdotal evidence suggests it is much more rewarding for a researcher's career to publish a paper with invalid results in a top economic journal than to publish a rigorous replication that overturns a result in a top journal.

What to do:

At the moment, the incentives are not present for editors or authors to invest a lot of effort in replications. Editors want to publish novel results since scientific recognition comes from innovative ideas, so replications are less rewarded and less welcome in leading journals. Even if published, a successful replication would validate the citation of the initial study. It would therefore typically not attract many citations on its own, unless a norm develops for replications to be cited along with the original study. Researchers therefore have more incentives to produce original research than to work on replications.

Changing these incentives is likely to be difficult, and any recommendation has to face the fact that proposals to change behaviour are unlikely to be effective when incentives are so clearly stacked against replication. Solutions have to consider changing the incentives. For instance: journals could consider having special issues dedicated to replications. It would allocate some predetermined space to replications without editorial decisions putting them in competition with non-replication studies, and a prize could be awarded to the best replication study. Replications can also make a natural training exercise for PhD students. Having avenues to publish replication studies could provide the needed incentive for them to do replications.[11] The size of replications is important, and the ESA may have a role in fostering multi-lab collaborations to replicate findings.

As mentioned, replications may often have an underwhelming effect on the ability of a research field to self correct by abandoning zombie results. Here again, improving this situation likely requires a change of incentives. One possible goal would be to decrease the lifetime of zombie results by improving the dissemination of information on failed replications.

A critical time when researchers need to be aware of existing replications is when they review the literature on a topic. When considering a specific reference, they would benefit from having quick access to the papers replicating the study. At the moment, finding replications is time consuming. Using Google Scholar, for instance, an author would have to browse many papers to find replications. Ideally, Google Scholar would include a link below each paper to "Replication/Reproduction" studies. Replication studies would reproduce the result with different researchers but the same experimental set up. One possible option could be to for a Replicability/Reproducibility repository platform (e.g. a wiki) to emerge. This platform could be filled by researchers themselves. Researchers running a replication would have the incentive to create a record for their target paper with a list of replications of the original result, including their own study.

[11] This idea was also suggested by Alex Roomets.

A public platform listing replications could also serve as a metric to judge the quality of a paper. Authors could then list their papers with their citations and the number of successful replication on their CVs. The idea of using replications as a metric of quality was suggested by Maniadis, Tufano and List (2015).

Maniadis, Tufano and List (2015) also considered the incentives of the original author to facilitate replication. They suggested an editorial policy of always allowing original authors to comment on replications. We envision that the replication and response would often occur in a journal other than the one in which the initial study was published. It has automatic benefits in terms of publications and citations, encouraging and rewarding discussions on research topics. The space required for comments can be small and therefore not be too costly for journals.

Another idea could be for the journals to publish, along with all the other stats they report, how many of their experimental studies have had new studies that replicate or have not replicated their results in the past year. For instance:

"22 new papers in 2021 that included replications of 17 papers published. Of these 22 papers, 20 replicated the original results while 2 failed to replicate the original results."

The journals could link to the list of replications. These would relevant give exposure to the replication studies and would contribute to increase the incentives of researchers to produce replications (as well as the incentives of editors to publish studies which are going to be replicated).

Summary and recommendations

The replication crisis raises important questions for the community of experimental economists. Several innovations to the process of publishing scientific studies have been proposed and they deserve the attention of the community of experimental economists. Our view on the different solutions proposed and the way to use them are summarised as follows:

Pre-registration:

They can be useful and it would be worth promoting them more. They should not be imposed as a blanket condition for publication in experimental economics. They should also not preclude the publication of exploratory analyses. Researchers should still be welcome to submit exploratory results, with possibly a more conservative way of appraising them. Journals could propose a standard format for published papers in which tests of hypotheses and exploratory analyses are reported in two different sections of the paper. With such solutions in mind, we believe that, if pre-registrations get more common, editors and reviewers should refrain from discouraging researchers from publishing exploratory results per se.

In order to help the reviewing process, while also limiting concerns about the theft of intellectual property, pre-registration can be private and only be available to reviewers/editors until publication of the research study. The pre-registration can be made public after the study is accepted for publication. The choice to make a pre-registration private could be left to the authors.

Registered reports:

Registered reports are an option worth considering. The discussions in the ESA around this document revealed some divergence of views about the merits of this approach. The ESA may

explore how its journals could allow such a publishing option in its journals, with a plan to assess its success.

There could be different types of registered reports: registration of hypotheses and design (like a high level pre-submission enquiry), registration of hypotheses, design and analysis plan (standard registered report). Initial editorial decisions could be less binding in the first case.

Replications:

Replications are highly desirable for scientific advancement, but currently do not provide enough of an incentive for most researchers. To induce more replications, changing incentives is therefore likely required. The visibility and citations of replications could be increased (via a public platform keeping track of original studies and their replications), and the chances of publication of replications could also be improved (via an increase in journal space).

To help foster replications, journals publishing experimental economics research could aim to have issues or sections dedicated to replications. Journals that publish experimental economics research could also include a link on the online version of the published paper to all future papers that include replications. The ESA could also encourage multi-lab collaboration to foster high quality replications. Authors whose studies are replicated could have a right to a comment on the replication published in the same issue, giving them an incentive to help replicating teams.

References

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y., 2016. Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *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. and Heikensten, E., 2016. Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), pp.1433-1436.

Coffman, L.C. and Niederle, M., 2015. Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, *29*(3), pp.81-98.

Dufwenberg, M. and Martinsson, P., 2014. Keeping researchers honest: The case for sealed-envelope-submissions. *IGIER (Innocenzo Gasparini Institute for Economic Research)*, (533).

Duyx, B., Urlings, M.J., Swaen, G.M., Bouter, L.M. and Zeegers, M.P., 2017. Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of clinical epidemiology*, *88*, pp.92-101.

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 *Replication in experimental economics*. Emerald Group Publishing Limited.

Simmons, J.P., Nelson, L.D. and Simonsohn, U., 2018. False-positive citations. *Perspectives on Psychological Science*, *13*(2), pp.255-259.