Editorial*

International Journal for Re-Views in Empirical Economics, Volume 2, 2018-2, DOI:10.18718/81781.6

Please Cite As: Grunow, Martina, Hilmar Schneider, Gert G. Wagner, and Joachim Wagner (2018). Editorial. *International Journal for Re-Views in Empirical Economics*, Vol 2(2018-2).
DOI: 10.18718/81781.6

Lack of reproducibility is seriously undermining the credibility of science as a whole. By extrapolating the findings of isolated checks, one may expect a substantial fraction of published articles in scientific journals to contain findings that simply do not hold. But we do not know which articles are affected - they usually remain uncovered.

Science is about improving our understanding of the world by formulating models and theories based on fundamental principles that consistently explain the emergence of observable phenomena. Empirical research is playing a key role for assessing the practical usefulness of our theories and helps in sorting out the relative power of competing theoretical explanations.
However, empirical research often comes with a certain lack of transparency, which makes it difficult for outsiders to assess the reliability of findings. It is hardly possible to reveal all the information necessary for comprehension of findings within the length of a typical article in a scientific journal. The fact that numerous findings are not replicable for whatever reason means that empirical research fails in fulfilling its role as an assessment device. This is the reason why more and more scientific journals are now requesting their authors to provide the datasets and the programming codes that have been used for achieving the statistical results. Unfortunately, this is still not enough. As impressively demonstrated in a number of papers presented during the 2017 annual meeting of the American Economic Association, even the fact that data and code are provided does not mean that the results of the related paper can be reproduced just on a technical level. And even in the case that results can be perfectly reproduced by re-running a given code over a given dataset that does not mean that the findings are reliable and robust. The results may be extremely sensitive to model specifications, the specification of certain variables or the definition of certain subsamples etc. This can only be found out by systematically varying specifications. Due to the widespread of, at least very often perceived, publication bias one should not expect too much here from the authors, because they may not want to lower their chances for publication by being too critical against themselves. Of course, this is a matter of culture within a scientific discipline. Thus, any means which help to establish a culture of replication are helpful.

Far beyond usual statistical confidence criteria

The problem we are talking about goes far beyond the usual statistical confidence criteria. Based on random samples, there is an inherent and inevitable risk of drawing false conclusions in any empirical study. A 5% level of statistical significance implies that on average five out of a hundred random sample draws might not be meaningful. In such cases we possibly draw erroneous conclusions that do not hold. This is known as the α error. There is also the risk of erroneously concluding to no effect, because an estimate lies within a certain confidence interval although there is an effect. This is known as the β error, which is much harder to assess in quantitative terms than the α error.
Based on available replication studies we can assume that roughly half of the findings published in academic journals are not replicable. This is ten times higher than one might expect from purely random deviations on a 5% significance level. In fact, if a statistically significant estimate is just a random deviation from a truly not existing effect, 95 out of 100 replications based on independent other random samples should yield non-significant estimates. If on the other hand there is a true non-zero effect, 95 out of 100 replications based on independent other random samples should yield significant estimates too, if the point estimate is identical to the true effect. If the true effect is smaller than the estimated effect, the replication ratio is likely to be smaller too.
On top, if replications are not based on another random sample draw but just technically reproduce the applied methodology based on the same sample (pure reproductions) the meaning of a deviation from the original estimation is clearly different from the meaning of random deviations.
Thus, an overall replication rate of approximately 50% suggests that we are somewhere in between α-error and not- random-non-replicable research. There are some statistically significant estimates that are just random deviates from no effect, there are some statistically significant estimates that are rightly pointing to a true effect but are over-estimating it, and there are some statistically significant estimates that are based on computational errors, misspecifications and the like as well as fraud. Replications can definitely help in sorting out things in a better way and help us in deepening our understanding of what is driving our world. Foregoing the chances of knowing better is generating social costs that might be huge.

Lack of credibility and urgent need for replications

There can be no doubt: systematic replications should be an integral part of the scientific process. But in fact, they are not. It is even worse: the scientific business as we know it is systematically evading scientific quality control by stigmatizing replications and discriminating against affirmations of already existing findings. The findings of researchers that investigated a phenomenon for the first time are dominating our scientifically substantiated picture of the world. Journals prefer to publish new and big effects so that original results are inflated: a bold effect - probably found just by chance - is much more likely to get published than a reliable small effect.
As a result, scientific journals are full of articles with findings that cannot be replicated. It is a central problem that in most cases no one knows which of these findings are replicable and which ones are not. This is not primarily a failure of the peer reviewing process. Peer reviewers are usually supposed to check for the scientific innovation of a paper, to assess the adequacy of the chosen theoretical and methodological approach and to assess its quality in light of what is regarded as state of the art in its specific respect. Detecting programming errors, uncovering sensitivities of findings with regard to various specifications is simply not the job of referees. If it were, it would become very difficult to recruit peer reviewers, because peer reviewing would become much more time-consuming as it already is. And it would extend the responsibility for possible errors to the

reviewers without the provision of an adequate reward for the additional effort.
To overcome this, one might consider a new type of peer reviewing process, consisting of independent institutions providing peer reviews on a professional base. However, this would require academics with a proven scientific reputation willing to devote a substantial part of their precious time to professional peer reviews instead of research. This implies a contradiction in itself. Therefore, we need to look for effective incentives to encourage replication studies.
Replications in a constructive manner do make a lot of sense for the scientific progress, but they are still rarely found in academic journals. The reasons for this are manifold. First of all, investing time into a replication is risky, because chances to publish the outcome in traditional journals are limited to those cases, where replicators happen to uncover really major flaws in a preceding paper. The chances to publish a replication, which mainly confirms more or less the mathematical correctness of previous findings, are close to zero. This publication bias does not take account of the importance of replicable results as a valuable guidance for future research to the scientific community.
Uncertainty about the possible outcome of a replication and the related risk of non-publication reduces incentives for replications and researchers may conclude that it might be better to invest time into something with higher chances for getting published. Moreover, replications are suffering from bad reputation inside academia. Researchers doing a replication are often regarded as lacking of inspiration.
By stigmatizing replications, science is effectively self-immunizing against possible critiques. This is nothing but a refusal to external quality control and a violation of the objectivity criterion, a fundamental principle of science.

Horizontal and vertical growth of knowledge

As a result, research - at least in economics but also in other disciplines - has degenerated to a process where knowledge is largely growing horizontally instead of vertically. Vertical growth of knowledge means that knowledge is constructively building upon previous findings according to the principle of Hegelian dialectics. Among else, this principle includes searching for explanations for contradictory findings (which is the idea of Popper, too). Replications and systematic variations from there are a natural way of extending knowledge in a constructive manner.
In contrast, horizontal growth of knowledge denotes a process where knowledge generation is mostly driven by the criterion of scientific innovation, which basically means doing something, which hasn't been done before. As a result, knowledge is emerging as a flat and unstructured mass of all kinds of ideas that do not sufficiently relate to each other. Findings supporting a specific hypothesis are peacefully resting besides findings supporting the opposite and no one cares. Scientific innovation as a key criterion for publications is preventing social and economic research from being relevant for society.

The International Journal for Re-Views in Empirical Economics (IREE)

By establishing the **I**nternational Journal for **R**e-Views in **E**mpirical **E**conomics (IREE), we want to make a contribution that may lead economic research back towards a discipline that is actively addressing contradictions as a major source of learning. We want to encourage authors to invest into investigations challenging previous findings, because this is a driver of scientific progress.
We welcome articles that are doing replications and re-views in the spirit of seeking for explanations of contradictions. To this end IREE publishes systematic reviews and articles dealing with

replication methods and the development of standards for replications. Complementary to this, IREE publishes descriptions of micro-datasets.
It is important to distinguish between different types of manuscripts.
Replications in our understanding are covering a large range starting from pure reproductions to scientific replications making use of variations over variable specifications, sample definitions, datasets, observational periods, countries and the like. By learning how such variations affect results we have reason to expect a much deeper understanding of social and economic phenomena. However, comparability to previous findings is key. By doing so, chances are that economics may be perceived as being relevant for society again.

Terra incognita

We are fully aware of the obstacles waiting for us along the way and we – editors as well as authors – are entering terra incognita in many respects. This is the reason why the section on methods and standards of replications is so vital for our journal. We will have to start with a number of preset conditions but they are likely to become adjusted with growing experience.

Confirmation, falsification, and publication biases

First of all, we will have to deal with the asymmetry between confirmation and falsification. While confirmation of previous findings is certainly less powerful with regard to assessing the reliability of findings than falsification, we do not want to discriminate against confirmative findings. In our view, this is a key condition to be set in order to get IREE going. Journals that have implemented a replication section often continue to publish only such replications that lead to "interesting" deviations from the original findings. By maintaining this kind of publication bias, it will not be possible to overcome the general reservations against replications as mentioned above. Publishing confirmative findings is not only important from a scientific point of view but it is also an important incentive: as authors can be sure that their replications will be published independent of the result and this minimizes the ex-ante risk of conducting and publishing replication studies (given that the replication is methodological sound). Our policy may lead to another "publication bias" towards confirmative findings in our journal, because falsifying results may also have chances to get published in other journals while confirmatory results may not. But this is nothing we worry about. For us it is important to make an effort towards increasing the number of replications in general.

Replicability criteria

Another question closely linked with the issue of confirmation and falsification is that of quantitative replicability criteria. Answers are depending on the type of replication. For pure reproductions of an article, this might just be a matter of pragmatism. With regard to scientific replications things are different. One might consider conventional inference measures of statistical differences. However, conceptual differences in the sampling technology might impose limitations that are likely to call for more sophisticated or simply pragmatic solutions. As discussed above, it is in the nature of random samples that the parameters we are interested in are only identifiable in a stochastic range. Repeated random sample draws are useful for narrowing down this range and may give rise to better point estimates. Nevertheless, how far away from the original figures a replication is allowed to be in order to be regarded as confirmed or failed should be formulated as a standard, which does

not exist yet. Therefore, we invite authors to propose such standards in our methodological section. As state of the art, manuscripts of pure reproductions submitted to IREE do not have to be long in terms of pages. They do not have to repeat the theoretical concept or other sections of a paper; they should just focus on the information that is relevant for comprehending the replication and explain the results. The manuscript of a scientific replication which is applying a certain method to a comparable dataset or a setting, which is different from the one in the original paper, should include a discussion of comparability of the method, data, results and conclusions with the original study. In any case, authors of replications must contact the authors of the replicated study and record this contact attempt before submitting their manuscript to IREE. The authors of the original study will have the opportunity to publish their comment along with the replication study.

Marginal value of replications

However, we see the need for limiting the number of pure reproductions of a specific article as the marginal value of a growing number of successful replications of a single article might rapidly decline to zero. Defining a maximum number is totally arbitrary, but three pure reproductions appear to be reasonable. In order to help avoiding useless effort, we suggest pre-registering for a planned replication with our journal. This pre-registration will be valid for a period of three months on a first come first serve base. It can be extended upon request for another three months, but only if no other pre-registration for this article is in the queue. The number of pre-registrations for certain replication attempts will be listed on our website. If already two pure reproductions of an article have been published in IREE, a third one will only be accepted for review, if the author is first in the pre-registration queue in case of pre-registrations.

Code of conduct and integrity

By establishing replications as a standard device of scientific quality control, there is a danger that it might become a habit to provide cheap confirmation service to friends, which in turn would undermine the general concept behind replications. In order to prevent this, it might require the development of a code of conduct to be signed by authors. However, for the time being, we count on the professional integrity of researchers.
Quality control is another issue to deal with. A pure reproduction that only consists of successful re-running provided code over a provided dataset is certainly of less power than a replication, where someone underwent the effort of re-writing programs and re-computing variables according to the verbal instructions given in a paper. The latter case is much better suited to discover programming errors or sensitivities to computational settings than the former. Confirmation based on re-computing should therefore be regarded as more valuable than confirmation based on simple re-running. But how can we distinguish the claim of having done a re-computation from having really done it? Here again, we have to rely on the professional integrity of researchers.

Citation standard for replications

Furthermore, we strongly intend to establish a common citation standard, which consists of citing the original article along with its replications. To the extent this is going to become a standard way of giving credit to published articles, both the original authors as well as their replicators will benefit as replications will increase citations of the replicated article itself and its replications.

It is our expectation that replications will receive citations to the same extent as original articles. From this point of view, doing replications and submitting them to IREE is likely to pay off as conventional citation counts will apply to IREE as to any other journal. If our expectations are right, IREE should tremendously benefit from citation counts, because it collects replications of articles of a whole set of journals instead of just one. In order to take benefit of this, IREE establishes a database of articles that have been replicated in IREE together with stylized results. We are confident to make it in a reasonable time to established citation indices.

Announcing a cultural revolution

With the founding of IREE, we are bold to announce a cultural revolution in the social sciences, which in our view is overdue. We are convinced that academia has to make such a move before tax payers will start raising questions about the usefulness of the social sciences. And we strongly hope not to underestimate the inertia of the established academic business. Evidence based research can be of invaluable importance for society if it is constructively oriented towards knowledge generation, and it lies in the responsibility of researchers to give credit to it.
We strongly feel that this is also the right point in time for starting such an initiative. Some disciplines, e. g. psychology, are already speaking of a replication crisis. The importance of the issue is documented in a growing number of articles in leading scientific journals calling for replications. The American Economic Association has put the papers of two replication sections very prominently in their proceedings of the annual meeting in 2017 published in the American Economic Review.

Acknowledgement

We are grateful for the fact that the German Research Foundation (DFG) and the ZBW Leibniz Information Centre for Economics are supporting our enterprise in the starting phase. We regard this as a strong indication for the relevance of what we have in mind.

Martina Grunow (Managing Editor)
Hilmar Schneider (Editor)
Gert G. Wagner (Editor)
Joachim Wagner (Chief Editor)