A partial solution for the replication crisis in economics

William M. Briggs
Statistician to the Stars!, Charlevoix, Michigan, USA

Received 25 March 2023
Revised 22 May 2023
Accepted 22 May 2023

Abstract

Purpose – Important research once thought unassailable has failed to replicate. Not just in economics, but in all science. The problem is therefore not in dispute nor are some of the causes, like low power, selective reporting, the file drawer effect, publicly unavailable data and so forth. Some partially worthy solutions have already been offered, like pre-registering hypotheses and data analysis plans.

Design/methodology/approach – This is a review paper on the replication crisis, which is by now very well known.

Findings – This study offers another partial solution, which is to remind researchers that correlation does not logically imply causation. The effect of this reminder is to eschew "significance" testing, whether in frequentist or Bayesian form (like Bayes factors) and to report models in predictive form, so that anybody can check the veracity of any model. In effect, all papers could undergo replication testing.

Originality/value – The author argues that this, or any solution, will never eliminate all errors.

Keywords Bayes factors, Hypothesis testing, Predictive modeling, Replication crisis

Paper type Viewpoint

1. Introduction

There is in economics, as there is in many other fields, a reproducibility crisis. Papers with results once thought sound, important, and even true are turning out to be unsound, unimportant, and even false, or at least not nearly as sure as thought.

Let us first briefly demonstrate the existence of the problem, which is by now fairly well known, review-proffered solutions, and then recommend our own partial solution, which begins with a review of our understanding of the philosophy of models, recognizing models are the lifeblood of economics – and of all of science. We end with a proposal to modify certain model efforts.

The replication crisis was recognized after several large efforts were made to reproduce well-known research. The efforts largely failed. Again, this is not just in one area, but anywhere models are used.

For example, Camerer *et al.* (2018) attempted to replicate 21 (in what most considered to be) important papers in the social sciences published in what many say are the best journals, *Nature* and *Science*.

In their replications, they used "sample sizes on average about five times higher than in the original studies." This means they did better than the originals, which had on average much smaller sample sizes. Even so, they only found "a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size."

Asian Journal of Economics and Banking
Vol. 7 No. 2, 2023
pp. 180-190
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-03-2023-0027

In other words, only about half the papers were found to replicate, and only at about half the effect sizes. This is a poor showing. This would seem astonishing, except it was discovered that this was in no way remarkable.

Take the situation in psychology. Klein *et al.* (2018) did the same as Camerer for 28 prominent psychology papers. This was a big effort. The replications "comprised 15,305 participants from 36 countries and territories."

Results: "Using the conventional criterion of statistical significance ($p < 0.05$), we found that 15 (54%) of the replications provided evidence of a statistically significant effect in the same direction as the original finding . . . Seven (25%) of the replications yielded effect sizes larger than the original ones, and 21 (75%) yielded effect sizes smaller than the original ones."

Again, only about half the papers replicated, with effect sizes mostly smaller than the original. As before, the papers chosen were considered to be important.

In medicine, Ioannidis (2005) examined 49 papers, all considered stellar efforts: each paper examined had garnered at least 1,000 citations each. Of the attempts at replication, "7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged." The story is the same as before. Only a quarter of the papers were left touched.

Richard Horton, editor of *The Lancet*, in 2015 announced that half of science is wrong (Horton, 2015). He said: "The case against science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness."

Open MKT tracks replication studies in marketing (Charlton, 2023). As of March 2023, only "5 out of 43 (11.6%) attempted replications in marketing were unambiguously successful." They also note that "The *Journal of Consumer Research* (Consumer Behavior's top journal) now stands at a 3.4% replication rate. So far, 29 replications have been attempted and one has succeeded." Just 1 out of 29 is dismal.

By "unambiguously successful" they mean "a significant result that matched the form of the original (same direction, same shape of the interaction effect, etc.). It also indicates that no obvious confounds appeared in the protocol when replicated."

Our concern here is economics. Alas, the outlook is just as bleak for this field. Camerer *et al.* (2016) discovered that only just over half of 18 famous experiments were replicated. These were, they claimed, the best papers under examination, and not the greater mass of ordinary research. They "found a significant effect in the same direction as in the original study for 11 replications (61%); on average, the replicated effect size is 66% of the original. The replicability rate varies between 67% and 78% for four additional replicability indicators, including a prediction market measure of peer beliefs."

It is easy to go on in this vein, the literature is by now quite large, and we do not pretend to have covered more than a small portion of it. Yet this is not necessary, because the theme is clear. There is an enormous problem with research. Error, falsities have been taken as scientific truths and over-certainty abounds. The question is: why and what can be done about it?

It must also be stressed that the half of science that is wrong, as Horton commented, and as all large replication efforts have confirmed, is the best science, or what is considered the best. Consider how bad it must be in the lower tiers of research, where work is far less prominent, less well-checked, and which exists in far greater number.

We next discuss some of the reasons the crisis exists and why so much over-certain science, marketing and economics research is produced.

2. Some reasons for the crisis

Smaldino and McElreath (2016) blame poor research design, which can lead to false-positive findings, but they emphasize that poor design, in concert with the very real need for researchers to publish to keep their jobs, creates a system in which bad science is "naturally selected." They also review the literature and find, to no one's surprise, that most studies are under-powered, which is to say, they have sample sizes much too small. That, in turn, makes it easier for false positives to flourish. See also Macleod and the University of Edinburgh Research Strategy Group (2022).

Baker (2016) writes of a survey of 1,576 researchers and their view of the reproducibility crisis. She says that "More than 70% of researchers have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments." This applies to fields as diverse as physics, chemistry, biology, medicine, earth sciences and of course economics. The problem is uniform and widespread.

Researchers in that survey, like with Smaldino and McElreath, blamed in part the "pressure to publish and selective reporting" for the crisis. Many noted researchers agreed that low statistical power is a problem. But they also mentioned factors like insufficient experimental guidance, inadequate detail in published methods, lack of publicly available data used in studies, poor experimental design, outright fraud, weak peer review and even bad luck. That last one is, for us, the most interesting, as we shall see in the next section.

Page *et al.* (2021) cite as the two main causes of the crisis the file-drawer effect, which is failing to research or publish negative or null findings, and P-hacking, the vigilant and all-too-often rewarding search for wee *p* values. See Bruns and Ioannidis (2016) about P-hacking and some tools that might be used to spot when it has occurred. It is far too easy to "generate" wee *p* values, which are mistakenly taken as "proving" true causal effects have been demonstrated. We say much more on this below.

To combat P-hacking, many suggest lowering the criterion of "significance," so that it is more difficult to find in any study. The precise value of the number after which "significance" is declared is now is so well known that we need not even write it. Can lowering it help? There are several difficulties with this idea. First, "significance" only means "*p* value less than some number". And a *p* value less than some number is "significant." This is circular. The problem, therefore, is the language itself. "Significant" does not mean important, true, useful or good. It especially does not mean true.

Even for those who see "significance" as a valuable tool, Williams (2019) shows that merely lowering the "significance" criterion can create even a greater number of false positives because of a negative selection effect.

Again, this is only a brief survey. But there is more than enough evidence that something, and of course likely many things, that have gone wrong with research. We next investigate some proposed solutions.

3. Some patches for the crisis

Page *et al.* (2021) suggest pre-registering trials, and especially pre-announcing trial hypotheses, along with publishing in advance the proposed analysis. Many authors echo this strategy. Indeed, this seems to be one of the most common solutions on offer. Banerjee *et al.* (2020) support their use in economics but warn their strict use is "detrimental to knowledge creation when implementing field experiments in the real world."

The idea of pre-registration is that when the data from a study finally comes in, the chance of P-hacking is lessened. There is far less freedom to "explore" hypotheses, which is the main drive of P-hacking. This should also cut down the file-drawer effect, at least to an extent, because researchers are always left free to modify their hypotheses *post hoc*. The idea is not foolproof because authors are still free to abandon their initial registrations and say, in effect, "although this was not our original goal, here is what we discovered."

Oddly, some others (see Fanelli, 2017) claim the chaos, false and over-certain results due to the crisis are “empowering”, and that if we removed the freedom to explore data, science would be harmed. We do not agree that accepting false results is “inspiring” or “empowering”.

Sharpe and Poets (2020) advocate meta-analysis to lessen the effects of the crisis. This is a reasonable approach, but it obviously only works in areas where there is already substantial literature. It cannot help the lone paper, or it cannot stop alone and mistaken paper from forming a following, in which more research along the lines of the mistaken paper is produced. Meta-analysis would only show agreement, not truth or accuracy.

Knuteson (2016) has the most amusing solution, perhaps the one most apt for economics research. He suggests scientists sell their research to journals. Publishers, in effect, bet on the accuracy of papers. “Each transaction is brokered by a central exchange, which holds money from the anonymous information buyer and anonymous information seller in escrow, and which enforces a set of incentives facilitating the transfer of useful, bluntly honest information from the seller to the buyer.” As odd as this sounds, it is not wildly different in spirit from what we propose below.

Trafimow (2018) suggests a method not too dissimilar from the one we advocate below as well. He believes if the researcher has well identified his probability model, the researcher should be able to “calculate the probability of replication, prior to data collection, and without dependence on the population effect size or expected population effect size.” This may well be so, but this calculation will be in error, as will be the researcher’s model, if the researcher has misidentified his population.

Miyakawa (2020) reminds us that fraud, in all its various forms, from self-deception to outright cheating, is a bigger problem than we might know. He was the editor of a prominent biology journal and asked 41 authors over three years to “revise” papers, requesting the authors provide raw data to the journal. Of the 41 articles, 21 were withdrawn, and “19 out of the remaining 20 manuscripts because of insufficient raw data.”

4. One (major) partial solution

While the suggestions in the previous section all have merits, they are often similar to medicine which treats the symptoms but not the source of a disease. The patient may feel better, and there is a certain palliative effect, but the illness has not gone away because it has not been treated. Far better, of course, to attack the cause of the disease, i.e. fix the problem at its base.

This is not always possible. Just as some diseases are invariably fatal, some are incurably chronic. That is the case with all science, economics included: perfection in our field cannot be had. There is no panacea, no evidentiary formula, which when applied assiduously, will lead inexorably to the truth. Mistakes will happen. That which is false will be called true, and vice versa. Bad models will always be with us.

Yet we might still attack the source of the problem itself, which is the philosophy of models. A solution which treats the disease, but offers only a partial and incomplete cure, is this: recall that correlation is not causation.

It is a staple, and truth, of logic that correlation does not imply causation. When a correlation is discovered, it may reveal causation, whole or partial, or it may reveal a merely coincidental occurrence. Though I have no proof of this, all common experience shows coincidence is the most common case in life. This is because almost any collection of data, no matter how unrelated, can be made to show correlations if it is examined closely enough. This is not interesting in itself – unless those correlations are claimed to be causation. This is what standard modeling procedures are designed to do.

There are three well-trodden paths or procedures in moving from observed correlation to claiming causation. Or if not directly claiming causation, then at least hinting at it strongly.

The first two apply to all statistical models: p values with null hypothesis significance testing (NHST), and Bayesian posteriors or Bayes factors. The third path is found in more fundamental, purely mathematical models, which build cause in the models directly.

4.1 Statistical models

Let us begin with statistical models. There are well-known differences in the philosophies of probability between frequentist and Bayesian theories, with Bayesian theory offering more than one interpretation, see Bernardo and Smith (2000), Briggs (2016) and Hájek (1996). We can pass by these differences and nuances and instead note the commonality of all statistical approaches, which is this: both frequentist and Bayesian methods include the tacit belief that "significant" probabilities, or large Bayes factors, indicate cause has been discovered or confirmed. And this is false, as I hope to show.

Both frequentist and Bayesian theories make great use of parameterized probability models. Some nonparametric models exist, but these are not different, for the most part, in what follows. One of the most common models is regression, which we will use here for our paradigm example, noting that our discussion applies to all parameterized models.

Regression begins with an observable y, and then moves to the usual *ad hoc* specification that uncertainty in y can be characterized by a parameterized normal distribution. There are, of course, myriad variants of this, but this example is familiar enough. The model is as follows:

$$y \sim N(\vartheta, \sigma^2), \tag{1}$$

where σ^2 is the spread (sometimes called the "variance") and ϑ is related to a series of measures x_i:

$$\vartheta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon. \tag{2}$$

NHST works by calculating a statistic relating to one or more of the β_j, and then calculating a p value of this statistic. If the p value is wee, the "null" hypothesis that $\beta_j = 0$ is rejected; see Briggs *et al.* (2019). This is equivalent to stating that $\beta_j \neq 0$. It is crucial to note that is an existential, causal statement. More on this in a moment.

Bayesian modeling begins in just the same way as in frequentism, with the same *ad hoc* model, but it includes "priors" on the values of the β_i and σ^2. These are later turned into "posteriors." The exact form of the posterior depends on both the *ad hoc* model and the prior. If this is, say, a conjugate prior, which specifies a normal for the β_i and an inverse gamma for σ^2, the posterior takes the form (see, e.g. Lancaster, 2004):

$$\rho(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto \rho(\boldsymbol{\beta} | \sigma^2, \mathbf{y}, \mathbf{X}) \rho(\sigma^2 | \mathbf{y}, \mathbf{X}), \tag{3}$$

where inference is usually on the density $\rho(\boldsymbol{\beta}|\sigma^2, \mathbf{y}, \mathbf{X})$. If, say, the probability $\beta_j > 0$ (or < 0) is greater than (1 - "significance number"), then, like with NHST, an existential statement is made, that β_j is "likely" not zero. That is the theory, at any rate. In practice, if $\Pr(|\beta_j| > 0|\sigma^2, \mathbf{y}, \mathbf{X})$ is high, it is decided, informally, that β_j is not zero. Of course, sometimes the posterior probability itself is not computed, owing to difficulties in machine calculation and the like, and Bayes factors are used instead. A Bayes factor is in abstract form this (or its inverse):

$$\frac{\Pr(D|\text{null model})}{\Pr(D|\text{full model})}, \tag{4}$$

where, in our example, the null model is the one with one or more $\beta_j = 0$, and where D is the data (all x and y). The Bayes factor functions like a p value, in that it is used to decide whether to set those θ_j with small Bayes factors to 0. There are some published guides giving an opinion about what size of Bayes factor is "significant" – though typically other words besides "significant" are given, but only to avoid clashing with frequentist lingo. However, none of these definitions is important to us, since regardless of which values are picked, the *logical status* of the model statements is the same.

Before discussing what these existential statements mean, let us consider a classic economic problem: a gambler betting at the roulette wheel. On an American version of the wheel, given that there are 38 slots, with 18 of them red, 18 black and 2 green, we deduce a probability model, which says that the chance of red is just under 0.474, and the same for black.

The gambler knows this model. He has, we suppose, just bet red four times in a row, yet black has come up each time. This frustrates him. The probability of this outcome is just about 0.05, which is the magic number to declare "significance" in statistical problems. This is the p value. According to frequentist procedure, the gambler should reject the "null" that the wheel is "fair." The Bayesian might or might not reject this "null," depending on the prior chosen for his model. But it is easy to see that if the Bayesian proceeds in the same manner as the frequentist, most common models and priors would agree with the frequentist answer. (If you don't like this example, increase the black run to 5.)

In any case, rejecting the "null" is an existential statement, it is saying the wheel is "unfair," which is also a causal statement, it describes something about the wheel itself. Rejecting the "null" declares there is something in the wheel, or in the wheel's environment, that is *causing* the wheel to deviate from the expectations of the "null" model. There is no escaping this *certain* causal language in frequentism, though a subjective Bayesian might be able to argue he is not certain. Yet even though he allows a measure of his uncertainty, that uncertainty still applies to a causal proposition.

The reader might rebel at the gambler's conclusion, and should, but he must realize that the gambler is right – *if* either frequentist or Bayesian theory is true. Both theories demand, on the pain of consistency, that the gambler must conclude a hidden cause exists or that one is highly certain.

There are many natural objections to the gambler's conclusion: this is not a big enough sample, this is not a good test, the p value should be set lower to be more demanding, the prior is high on "fair", and so on. These are natural and good objections but notice they all are invoked because we recognize the absurdity of the NHST or Bayesian solution. Everybody knows the gambler has fallen prey to the well-known gambler's fallacy. And we know this because we know, or have every reason to believe, nothing has interfered with the causes of the wheel, or of the wheel's environment.

At first, a more serious objection is that this problem is too simple and that we should not use parameterized models and their formal theories in situations like this. Yet if that is a good objection, then the following question becomes obvious, and indeed mandatory: how do you know when to use hypothesis testing, frequentist or Bayes?

The answer is: *you do not.* There is nothing in the probability theory of either frequentism or Bayes that says "use testing here, but not here." It's all or nothing. If testing works as advertised, it should work everywhere. If it fails, however often, and as it did here, then it must not ever be used to infer cause because it is never known, in advance, whether the usage *this time* is appropriate or not. This conclusion is damning or should be.

This is the answer, though it is, or will be, unpalatable. Too many researchers believe they can pick and choose those times they invoke their theory of probability and testing, and when not, discarding testing when it does not seem appropriate, using it when the spirit takes them. Testing becomes a tool in a toolkit, and the philosophy which gave testing its justification is forgotten and discarded. Yet if that philosophy is true, or at least if it is believed, it must be embraced every time, not just sometimes. Because testing in practice is used at whim, and not consistently, as would be absurd anyway, we are forced to conclude that testing cannot, and should not, be trusted.

Testing, after all, is what gave all (or many of) those wrong answers in the crisis; testing is what caused papers to fail to replicate.

Frequentist theory says that if the test level of, say, 0.05 is chosen, then mistakes will happen 1 time in 20. That means one bad paper in 20 will be published. That one bad paper will make a causal claim that turns out false. Yet, as we have seen, the actual error is around 50%, or even much higher, especially if we consider all published papers and not just the "best" ones tested in replication studies.

If we knew the causes of the observable under study in advance, like the wheel or in any of all the problems to which statistics is put, which is to say, all the many thousands of models created and published every day, then we would not need to test and would not. This point must be appreciated fully.

The gambler has witnessed a coincidence, and, misled by his statistics lessons, committed a fallacy. The researcher who announces his wee p value or large Bayes factor does the same. Every time a p value or Bayes factor is used to make a decision that cause exists, a formal fallacy has been committed. This is so even if the correlation is indeed a cause. What I mean is that the conclusion "cause exists" does not logically follow from the premises "a correlation has been observed" and "a wee P (or large Bayes factor) has been seen." This is an inescapable fact of logic. Importantly, a researcher's "confidence" (I'm using this word in its plain English connotation) that he has done the right thing has no bearing on logic.

The partial solution to the replication crisis, then, is to eschew testing, in any of its forms, frequentist, non-parametric or Bayesian, and to publish models in a form which emphasizes model uncertainty. This can be done by putting models in their predictive form – and by recalling *all* models that can be put in a predictive form. Here is a schema of the form (the reader can see easily how this follows in the regression example):

$$\Pr(y \in s|X, \mathbf{y}, \mathbf{X}, \mathrm{M}), \tag{5}$$

where $\mathbf{y}$, $\mathbf{X}$ are the previous observations, i.e. the data D, and where X is a projection or guess what of future values of the measure X will take. This gives the probability y will be in the set s. That set is chosen to be "interesting" to the audience most likely to use the model. Finally, M is the model advocated by the researchers. In Bayesian terms, this is the posterior predictive form of the model; in frequentist terms, it is just the predictive form.

Of course, more than one prediction of $y \in s$ can be made, not only for multiple s, but for as many models M_i as the authors care to the state. In this way, anybody who can make a guess about X can check the model for themselves. More about the predictive form of statistical models is found in Briggs (2019), Briggs and Nguyen (2019), Bose (2004), Clarke and Clarke (2018) and Smith (1998).

Stating models in the predictive form will not stop authors from making causal claims. After all, if $\Pr(y \in s|X, \mathbf{y}, \mathbf{X}, \mathrm{M}_{\mathrm{null}})$ is very different from $\Pr(y \in s|X, \mathbf{y}, \mathbf{X}, \mathrm{M}_{\mathrm{full}})$, the temptation to make a causal claim will be overwhelming. That is one reason this solution is only partial. It does not eliminate untoward causal claims. Nothing can. Not if the author is determined to make one.

What this partial solution does, though, is to ensure researchers know how easy it is for anybody, everybody, to check their results. Critics do not have to have access to researcher's data, they do not need to know whether researchers engaged in *post hoc* reasoning, they do not need to see any computer code, they don't even have to know if researchers cheated or fooled themselves. All critics have to do is to check whether the claim that $\Pr(y \in s|X, \mathbf{y}, \mathbf{X}, \mathrm{M})$ is good or not. Nothing could be simpler for those tasked with checking model goodness.

This realization that their model is exposed to the world, and to all forms of open criticism, will, it is hoped, lead to at least a fraction more caution and circumspection from researchers. They must take great care or they will suffer at least the scorn of their colleagues for publishing a bad model.

4.2 Causal models

This brings us to the second form of economic models, which are directly mathematical or theoretical, and all use open causal statements in their theories. They may use statistics in these models, indeed they do as observation, but they are not purely probability-based models or theories. As an example, I'm thinking of such things as Keynesian vs Classical economic theories.

For instance, Keynesian theory declares that in a recession, governments should borrow (or deficit spend) and use the money to make up for missing public spending. Whereas Classical theory declares government should always maintain a balanced budget. Both theories make direct causal claims and are put in plain and, at least not always, in obscure probabilistic language.

Which theory is true, Keynesian or Classical, or indeed some other is not important here. What is important is that even though we have had about a century's worth of observation, both theories still have proponents, and both still have critics. Neither side has declared defeat, though both have announced victory, and everybody in all camps is adamant.

How does this happen? By careful choice of data and circumstance with which the models' pronouncements are validated or rather checked against. It is very easy to pick and choose examples which support one's favored theory. This results in the old joke that if you ask two economists their views, you'll get three opinions. Or more.

The point, which is anyway obvious, is that being able to openly check a model is not a be-all, end-all solution. There are still major and serious disagreements with just about every economic theory ever proposed. Even though everybody has been checking, and rechecking, these models, and for a very long time.

The same will be true when purely statistical models are put in predictive form. It will be almost impossible to conclusively falsify any of them. Therefore, the only real benefit of this partial solution is to increase the stated uncertainty in models – which is equivalent to decreasing the commonly overstated over-certainty – and this is surely salutary.

5. Conclusion

The number of replication studies is growing, according to Reed (2017). These range from only single digits in the late 1990s to over 80 from 2012 to 2016. Others call for more regular, even programmatic, replication efforts; see Harvey (2019). These efforts help will alleviate the crisis, but, in the end, like peer review, they can easily lead to a further bureaucratization of publishing.

In what at first might be thought to be good news, Mede *et al.* (2021) find that the public is for the most part unaware of there is a replication crisis. They discovered that once some members of the public are informed there is a crisis, most put it down to experimental errors being inevitable in science. They don't worry about it over much, because they consider that science is "self-correcting."

Other members of the public are, however, aware of the crisis. Mede points out the obvious implication: that since there is a crisis, science is given too much trust. Because more errors are publicly exposed, the public is increasingly learning to distrust all of science, and of course, economics has never had universal trust. Since it is never known in advance which science is good and which bad, you would think it would lead to greater humility among researchers when they communicate their results. Alas, this is not so. Hossenfelder (2017) points us to an extreme case of hubris in physics, a field which is increasingly giving up empirical verification of the predictive approach, and embracing model beauty as a measure of success.

There is in the literature support for the use of classical hypothesis testing, even in the face of the replication crisis, see Spanos (2022). But this support does not address the ideas advocated here, which is to put models in checkable, predictive forms. Instead, the suggestion is to ask authors to, in effect, be more careful in testing. Even if researchers are as careful as they can be, they cannot avoid making the fallacy that correlation is not causation. And, as the references above about the predictive method show, parameter-based methods invariably are over-certain because they put results in terms of unobservable parameters, and not in terms of observables themselves.

Having researchers put their models in predictive terms will allow these models to be checked by anybody, at least theoretically. This in effect allows all models to be replicated, in a sense. All of the research would be replicable, in the sense that all models and theories can be openly verified. Further, since models will be put in terms of observables, and not parameters, it will be much harder to overstate the effect size of models. This follows logically from the deduction that one can know a parameter or parameters with absolute certainty, or have a *p* value identically equal to 0, but the uncertainty in the observable can still be vast. It is odd that this simple truth is so often forgotten.

In case this truth has been forgotten by the reader, suppose the uncertainty in amounts of the reserve for two banks in successive measurements is characterized (modeled) by a normal distribution with central parameters ϑ_1 and $\vartheta_2 = \vartheta_1 + 1$, in dollars (with the same spread parameters). We are *certain*, the probability is 1, that $\vartheta_2 > \vartheta_1$, but the difference is a trivial, and less than trivial, one dollar. The certainty in the parameters does *not* translate to certainty in the reserves themselves/

We thus must eliminate "statistical significance", the obsession with unobservable parameters and state research in terms of *practical* significance.

In this way, it will be harder to overstate the importance of research, but of course, it will not be impossible. As with purely causal models, like Keynesian vs Classical economics, evidence can always be found by clever researchers that justify beliefs in theories. All experience proves this. This means that there is no overall solution to the replication crisis since mistakes can always be made. The best we can hope for is to reduce the chance of error.

References

Baker, M. (2016), "1,500 scientists lift the lid on reproducibility", *Nature*, Vol. 533, pp. 452-454.

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L.F., Olken, B.A. and Sautmann, A. (2020), *In Praise of Moderation: Suggestions for the Scope and Use of Pre-analysis Plans for Rcts in Economics*, Working Paper 26993, National Bureau of Economic Research.

Bernardo, J.M. and Smith, A.F.M. (2000), *Bayesian Theory*, Wiley, New York.

Bose, S. (2004), "On the robustness of the predictive distribution for sampling from finite populations", *Statistics and Probability Letters*, Vol. 69 No. 1, pp. 21-27.

Briggs, W.M. (2016), *Uncertainty: The Soul of Probability, Modeling and Statistics*, Springer, New York.

Briggs, W.M. (2019), "Reality-based probability and statistics: solving the evidential crisis", *Asian Journal of Business and Economics*, Vol. 3 No. 1, pp. 37-80.

Briggs, W.M. and Nguyen, H.T. (2019), "Clarifying asa's views on p values in hypothesis testing", *Asian Journal of Business and Economics*, Vol. 3 No. 2, pp. 1-16.

Briggs, W.M. (2019), "Everything wrong with p values under one roof", in Kreinovich, V., Thach, N.N., Trung, N.D. and Thanh, D.V. (Eds), *Beyond Traditional Probabilistic Methods in Economics*, Springer, New York, pp. 22-44.

Bruns, S.B. and Ioannidis, J.P.A. (2016), "p-curve and p-hacking in observational research", *PLoS ONE*, Vol. 11 No. 2, e0149144.

Camerer, C.F., Dreber, A., Forsell, E., Ho, T.-H., Huberand, J., Johannesson, M., Kirchler, M., Almenberg, J. and Altmejd, A. (2016), "Evaluating replicability of laboratory experiments in economics", *Science*, Vol. 351 No. 6280, pp. 1433-1436.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J. and

Wu, H. (2018), "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015", *Nature Human Behaviour*, Vol. 2 No. 9, pp. 637-644.

Charlton, A. (2023), "Replications of marketing studiesass", available at: https://openmkt.org/research/replications-of-marketing-studies/

Clarke, B.S. and Clarke, J.L. (2018), *Predictive Statistics*, Cambridge University Press, Cambridge.

Fanelli, D. (2017), "Is science really facing a reproducibility crisis, and do we need it to?", *PNAS*, Vol. 115 No. 11, pp. 2628-2631.

Hájek, A. (1996), "Mises redux—redux: fifteen arguments against finite frequentism", *Erkenntnis*, Vol. 45 Nos 2-3, pp. 209-227.

Harvey, C.R. (2019), "Replication in financial economics", *SSRN*, available at SSRN: https://ssrn.com/abstract=3409466

Horton, R. (2015), "Offline: what is medicine's 5 sigma?", *The Lancet*, Vol. 385 No. 9976, p. 1380.

Hossenfelder, S. (2017), "Science needs reason to be trusted", *Nature Physics*, Vol. 13 No. 4, pp. 316-317.

Ioannidis, J.P.A. (2005), "Contradicted and initially stronger effects in highly cited clinical research", *JAMA*, Vol. 294 No. 2, pp. 218-228.

Klein, R.A., Vianello, M., Hasselman, F., Adams, B.G., Adams, R.B. Jr, Alper, S., Aveyard, M., Axt, J.R., Babalola, M.T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M.J., Berry, D.R., Bialobrzeska, O., Binan, E.D., Bocian, K., Brandt, M.J., Busching, R., Rédei, A.C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C.L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E.E., Cheong, W., Cicero, D.C., Coen, S., Coleman, J.A., Collisson, B., Conway, M.A., Corker, K.S., Curran, P.G., Cushman, F., Dagona, Z.K., Dalgar, I., Rosa, A.D., Davis, W.E., de Bruijn, M., De Schutter, L., Devos, T., de Vries, M., Doğulu, C., Dozo, N., Dukes, K.N., Dunham, Y., Durrheim, K., Ebersole, C.R., Edlund, J.E., Eller, A., English, A.S., Finck, C., Frankowska, N., Freyre, M.Á., Friedman, M., Galliani, E.M., Gandi, J.C., Ghoshal, T., Giessner, S.R., Gill, T., Gnambs, T., Gómez, Á., González, R., Graham, J., Grahe, J.E., Grahek, I., Green, E.G.T., Hai, K., Haigh, M., Haines, E.L., Hall, M.P., Heffernan, M.E., Hicks, J.A., Houdek, P., Huntsinger, J.R., Huynh, H.P., Ijzerman, H., Inbar, Y., Innes-Ker, Å.H., Jiménez-Leal, W., John, M.-S., Joy-Gaba, J.A., Kamiloğlu, R.G., Kappes, H.B., Karabati, S., Karick, H., Keller, V.N., Kende, A., Kervyn, N., Knežević, G., Kovacs, C., Krueger, L.E., Kurapov, G., Kurtz, J., Lakens, D., Lazarević, L.B., Levitan, C.A., Lewis, N.A., Jr, Lins, S., Lipsey, N.P., Losee, J.E., Maassen, E., Maitner, A.T., Malingumu, W., Mallett, R.K., Marotta, S.A., Mededovic, J., Mena-Pacheco, F., Milfont, T.L., Morris, W.L., Murphy, S.C., Myachykov, A., Neave, N., Neijenhuijs, K., Nelson, A.J., Neto, F., Nichols, A.L., Ocampo, A., O'Donnell, S.L., Oikawa, H., Oikawa, M., Ong, E., Orosz, G., Osowiecka, M., Packard, G., Pérez-Sánchez, R., Petrović, B., Pilati, R., Pinter, B., Podesta, L., Pogge, G., Pollmann, M.M.H., Rutchick, A.M., Saavedra, P., Saeri, A.K., Salomon, E., Schmidt, K., Schönbrodt, F.D., Sekerdej, M.B., Sirlopú, D., Skorinko, J.L.M., Smith, M.A., Smith-Castro, V., Smolders, K.C.H.J., Sobkow, A., Sowden, W., Spachtholz, P., Srivastava, M., Steiner, T.G., Stouten, J., Street, C.N.H., Sundfelt, O.K., Szeto, S., Szumowska, E., Tang, A.C.W., Tanzer, N., Tear, M.J., Theriault, J., Thomae, M., Torres, D., Traczyk, J., Tybur, J.M., Ujhelyi, A., van Aert, R.C.M., van Assen, M.A.L.M., van der Hulst, M., van Lange, P.A.M., van 't Veer, A.E., Vásquez-Echeverría, A., Vaughn, L.A., Vázquez, A., Vega, L.D., Verniers, C., Verschoor, M., Voermans, I.P.J., Vranka, M.A., Welch, C., Wichman, A.L., Williams, L.A., Wood, M., Woodzicka, J.A., Wronska, M.K., Young, L., Zelenski, J.M., Zhijia, Z. and Nosek, B.A. (2018), "Many labs 2: investigating variation in replicability across samples and settings", *Advances in Methods and Practices in Psychological Science*, Vol. 1 No. 4, pp. 443-490.

Knuteson, B. (2016), "The solution to science's replication crisis".

Lancaster, T. (2004), *An Introduction to Modern Bayesian Econometrics*, Blackwell Publishing, New York.

Macleod, M. and the University of Edinburgh Research Strategy Group (2022), "Improving the reproducibility and integrity of research: what can different stakeholders contribute?", *BMC Research Notes*, Vol. 15 No. 1, p. 146.

Mede, N.G., Schäfer, M.S., Ziegler, R. and Weißkopf, M. (2021), "The "replication crisis" in the public eye: Germans' awareness and perceptions of the (ir)reproducibility of scientific research", *Public Understanding of Science*, Vol. 30 No. 1, pp. 91-102, PMID: 32924865.

Miyakawa, T. (2020), "No raw data, no science: another possible source of the reproducibility crisis", *Molecular Brain*, Vol. 13, pp. 1-6.

Page, L., Noussair, C.N. and Slonim, R. (2021), "The replication crisis, the rise of new research practices and what it means for experimental economics", *Journal of the Economic Science Association*, Vol. 7 No. 2, pp. 210-225.

Reed, W. (2017), "Replication in labor economics", *IZA World of Labor*, Vol. 413.

Sharpe, D. and Poets, S. (2020), "Meta-analysis as a response to the replication crisis", *Psychologie canadienne*, Vol. 61 No. 4, pp. 377-387.

Smaldino, P.E. and McElreath, R. (2016), "The natural selection of bad science", *Royal Society Open Science*, Vol. 3 No. 9, doi: 10.1098/rsos.160384.

Smith, R.L. (1998), "Bayesian and frequentist approaches to parametric predictive inference (with discussion)", *Bayesian Statistics*, Oxford University Press, Vol. 6, pp. 589-612.

Spanos, A. (2022), "Frequentist model-based statistical induction and the replication crisis", *Journal of Quantitative Economics*, Vol. 20 No. 1, pp. 133-159.

Trafimow, D. (2018), "An a priori solution to the replication crisis", *Philosophical Psychology*, Vol. 31 No. 8, pp. 1188-1214.

Williams, C.R. (2019), "How redefining statistical significance can worsen the replication crisis", *Economics Letters*, Vol. 181, pp. 65-69.

Corresponding author
William M. Briggs can be contacted at: matt@wmbriggs.com