**A Mistaken Confidence in Data**

**Edouard Machery**

**University of Pittsburgh**

**Machery@pitt.edu**

**Abstract**

In this paper I explore an underdiscussed factor contributing to the replication crisis: Scientists, and following them policy makers, often neglect sources of errors in the production and interpretation of data and thus overestimate what can be learnt from them. This neglect leads scientists to conduct experiments that are insufficiently informative and science consumers, including other scientists, to put too much weight on experimental results. The former leads to fragile empirical literatures, the latter to surprise and disappointment when the fragility of the empirical basis of some disciplines is revealed.

**1. Introduction**

The replication crisis raises various issues: Is there a crisis at all (e.g., Fanelli, 2018)? If there is a crisis, what factors contributed to its emergence (e.g., Nelson, Simmons, and Simonsohn, 2018)? And what can be done about it (e.g., Nosek, Spies, and Motyl, 2012)? In this article, I will take for granted that there is a replication crisis in several scientific disciplines (in a sense clarified in Section 1 below), and I will focus on the factors that contribute to this crisis.

A common explanation is that the replication crisis is due to questionable research practices, including both publication bias and p-hacking (the practices increasing the probability of obtaining a significant result), which result in empirical literatures littered with false positives (Simmons, Nelson, and Simonsohn, 2011). Questionable research practices are without any doubt part of the explanation, but in this paper I explore another factor contributing to the replication crisis: a mistaken confidence in data. The basic idea, to be elaborated at greater length in what follows, is that scientists, and following them lay people, including policy makers, often neglect sources of errors in the production and interpretation of data and thus overestimate what can be learnt from them. This neglect leads scientists to conduct experiments that are insufficiently informative and science consumers, including other scientists, to put too much weight on experimental results. The former leads to fragile empirical literatures, the latter to surprise and disappointment when the fragility of the empirical basis of some disciplines is revealed. In this article I will examine two important forms this mistaken confidence in data can take: neglecting sampling error and neglecting measurement error. I mostly focus on psychology, but the points made here can be extended to other scientific disciplines, and indeed I will give examples from the biomedical and other sciences.

Examining whether the replication crisis results in part from a neglect of some sources of error is in line with the error-theoretic approach in the philosophy of science (e.g., Hon, 1989; Mayo, 1996; Allchin, 1996). Error-theoretic philosophers of science have compellingly argued that probing for errors and correcting them is a crucial source of science’s trustworthiness. This idea would be supported if the neglect of sampling and measurement errors contributed to the replication crisis.

Here is how I will proceed: I briefly review the replication crisis in Section 2; in Section 3, I turn to the neglect of sampling error and its role in the replication crisis; in Section 4, I examine the neglect of measurement error and its role in the replication crisis. I conclude by emphasizing the alluding, but misleading appearance of certainty that often surrounds data.

**2. The Replication Crisis**

*2.1 The Fragility of Empirical Literatures*

In many scientific disciplines, the last decade has been marked by the realization that empirical literatures are much more fragile than previously thought, in the sense that many empirical results in these literatures cannot be successfully replicated (because they are false positives, because of the heterogeneity of the populations data are sampled from, or for whatever other reason).

Psychology has been at the forefront of this realization. Most famously, Nosek and colleagues attempted to replicate 100 psychological experiments drawn from a leading journal in cognitive psychology and two top journals in social psychology (Open Science Collaboration, 2015). While there are several criteria for deciding whether a replication succeeds, these criteria converged to suggest a low replication rate: Only 36.1% successfully replicated, as measured by the number of significant results in the replications reported in Open Science Collaboration (2015), and 41% as measured by the proportion of 95% confidence intervals of the effect sizes in the replications that included the original effect sizes. While it is not possible to estimate the rate of false positives in psychology from the Open Science Collaboration’s findings, at the very least we can confidently say that a surprisingly large number of results in social and cognitive psychology are unlikely to replicate and that some of them are likely to be false positives. This conclusion is not limited to psychology. The same pessimism seems to apply to the biomedical sciences, including oncology (e.g., Begley & Ellis, 2012), ecology (Lemoine et al., 2016), and economics (Chang & Li, 2015).

 Admittedly, the significance of this and similar findings has been extensively debated over the last few years, some objecting to the alarmist conclusion that there is a crisis in psychology and other scientific disciplines. For the sake of space, I will not engage with this debate in detail here, but the following remarks may be useful. First, it is important to distinguish two claims: the claim that many empirical literatures are fragile in the sense specified above and the claim that this fragility produces a crisis in psychology, or perhaps in science in general. Some have denied that empirical literatures in psychology and other sciences are fragile, but I take this position to be indefensible after a decade of reports of failed replications on a weekly basis. More plausibly, others grant that some empirical literatures are fragile, but downplay the significance of this fragility, either because they highlight the importance of errors in the normal course of science or because they take non-replicable results to have a negligible influence on the course of science (e.g., Fanelli, 2018). There is no doubt that errors are common in the history of science and are arguably part of its normal course, but this does not mean that all types of errors are harmless or can contribute to progress. In addition, the doubts that are now overshadowing the last four decades of social psychology as well as contemporary research in this field show that the fragility of an empirical literature can be extremely costly.

*2.2 Ghost Literatures*

The fragility of scientific literatures is not the only important characteristic of the replication crisis. The second important characteristic is the existence of what I will call “ghost literatures”: Ghost literatures are empirical literatures with numerous findings providing apparent convergent support for the reality of a phenomenon that is in fact not real. The extensive literatures on power pose and on ego depletion are two ghost literatures (e.g., Hagger et al., 2016; Simmons & Simonsohn, 2017). Ghost literatures can range from a few dozens articles, as was the case in the literature on power pose when it went down crashing, to several hundreds of articles, as is the case of the literature on ego depletion. Some ghost literatures are at forefront of science: They represent one of the on-going developments of a given science. The literature about power pose is an example. Others have been assimilated into the empirical core of science—the set of findings that scientists take to be established—as was the case of the literature about ego depletion.

Ghost literatures are rarely made up of so-called exact replications (roughly, a replication is exact if and only if it is similar to the original experiment except for its sample of participants; for discussion see Machery, 2020); rather they typically involve a large number of conceptual replications (roughly, conceptual replications modify the measurement or manipulation of an original experiment). Other articles in ghost literatures examine moderators and thus claim to identify boundary conditions of effects; others focus on mediators, often aiming at explaining the reported phenomena; yet others connect the findings with other theories or literatures.

 Ghost literatures show not only patterns of convergence, but also patterns of accumulation: Findings apparently build on previous findings and theories are elaborated on the basis of previous theories. Ghost literatures appear cumulative because moderators and mediators are typically a focus of scientists contributing to them. Scientists identify populations in which the phenomenon of interest (e.g., power pose) is not observed, or where it is observed more weakly; they identify variables that seem to be causally involved in the production of the phenomenon of interest. For instance, many mechanistic theories have been proposed to explain ego depletion: These disagree about which variables are assumed to explain why past self-control depletes will.

 One may ask how common ghost literatures are in psychology and other sciences. It is difficult, maybe impossible, to answer this question with confidence since, just like fraud, ghost literatures remain invisible until being revealed. They are however not extremely rare since it not exceedingly difficult to list other ghost literatures in psychology, psychiatry, neuroscience, and the biomedical sciences: social priming in social psychology, the penetrability of perception by desires and beliefs (i.e., the new look and its regular revivals—Machery, 2015), the influence of incidental emotions such as disgust on moral judgment, the cognitive benefits of bilingualism, the contribution of 5-HTTLPR to psychiatric conditions, (perhaps controversially) the serotonin hypothesis of depression, the significance of unsaturated fat for health, and the role of oxytocin in social interactions.

 To summarize, the replication crisis has two main components: the fragility of empirical literatures, with a large proportion of unreplicable empirical results, and the existence of ghost literatures.

**3. The Neglect of Sampling Error**

*3.1 Sampling Error*

Data tell us something about the state of the world (e.g. a given population parameter or a given physical constant), but the information they provide is more or less uncertain. The point of this paper is that scientists routinely underestimate this uncertainty, and that this underestimation is one of the sources of the replication crisis. The point of this specific section is to focus on a first source of this underestimation—the neglect of sampling error—and how it leads to fragile empirical literatures.

 Sampling error is simply the distance between the sample and the population parameters, for instance the means of a sample and of the population this mean is drawn from. Sampling error explains why two samples are often different, and why samples are not necessarily made at the image of the relevant populations. As the sample increases in size, the less likely it becomes that the sampling error exceeds a given threshold. Someone neglects sampling error just to the extent that she underestimatew how large sampling error is when she designs or interprets experiments: As a result, she overestimates how much information small sample sizes provide about the values of parameters and underestimates how large sample sizes must be to provide satisfactory information about these values.

The neglect of sampling error has epistemic and volitional sources. From an epistemic point of view, someone has a poor grasp of sampling error if she does not understand the impact of sampling error on hypothesis testing or parameter estimation; she may know that sampling error matters while failing to appreciate how much it matters (on the failure of statistical education in the behavioral and biomedical sciences, see, e.g., Windish et al., 2007). Someone might also fail to act on her grasp of sampling error: Her problem is then volitional. A scientist may be incentivized to run small studies, for instance (e.g., Higginson & Munafò, 2016), and she may rationalize her suboptimal decisions about experimental design by appealing to the need to make compromises. In practice the neglect of sampling error is often both epistemic and volitional (although there might also be variation across people): Scientists are rarely fully aware of the impact of sampling error, and they are incentivized to obfuscate the issue.

I distinguished the neglect of sampling error and, to anticipate, measurement error from questionable research practices in the introduction of this article, but one might wonder whether this neglect isn’t just another questionable research practice. The notion of questionable research practices is very vaguely defined. If questionable research practices include all practices that tend to undermine the trustworthiness of science while not clearly violating community-wide norms (in contrast to fraud and misconduct), then the neglects of sampling and measurement errors count as questionable research practices; if the notion of questionable research practices is more narrowly defined to include only those practices that are normatively dubious, although not clearly counternormative, the neglect of sampling and measurement error may not count as a questionable research practice since at least until recently, and perhaps still currently, there was no norm against experimental designs neglecting these two forms of error: For instance, the typical justification for sample size choice was, and to some extent still is, conformity with past practices. Be it as it may, most characterize questionable research practices extensively (i.e., as including publication bias and p-hacking), and, so defined, they do not include the neglect of sampling and measurement error.

*3.2 The Belief in the Law of Small Numbers*

People do not intuitively grasp how different a sample can be from the population from which it is drawn and how large a sample has to be in order to blunt the impact of sampling error. Let’s suppose you want to test the fairness of a coin, which happens to be in fact slightly biased (10%): In the long run it falls on head 4 times out of ten. How often should you throw that coin so as to reach a .8 probability to observe a number of heads or tails such that the probability of observing that number or a larger number if the coin is fair is lower than .05? Or, as psychologists usually put it, so as to have a power equal to .8 to reject the null hypothesis that the tail is fair at the .05 significance level? Write down your answer and keep reading. If you wrote 20, you are way off, in fact off by about an order of magnitude; if you wrote 50, you are still off by a factor of nearly 4; if you wrote, 100, you are closer, but still way off. The correct answer is 194 throws: It is only if you throw the coin 194 times that the probability of rejecting the hypothesis of a fair coin reaches .8, for a significance level of .05 and a 10% bias. Figure 1 presents the power curve for this test.



Figure 1: Power Curve for a Two-Sided Binomial Proportion Test With a 10% Bias.

The lesson to draw from this simple example is that people tend to underestimate how many data points are needed to be able to estimate a given population parameter to a given degree of precision or to reject a hypothesis correctly at a given level of significance, and relatedly massively overestimate the probability of getting a correct answer with a small sample.

Do scientists fare any better? Psychologists don’t have a good grasp of the uncertainty involved in data and how much data would be needed to counteract this uncertainty, or if they do they don’t seem to pay heed to it. In a classic article (1971), Tversky and Kahneman gave psychologists the following survey:

Suppose one of your doctoral students has completed a difficult and time-consuming experiment on 40 animals. He has scored and analyzed a large number of variables. His results are generally inconclusive, but one before-after comparison yields a highly significant t = 2.70, which is surprising and could be of major theoretical significance. Considering the importance of the result, its surprisal value, and the number of analyses that your student has performed— Would you recommend that he replicate the study before publishing? If you recommend replication, how many animals would you urge him to run?

Most psychologists recommended replication, and estimated on average that a sample size of 20 participants would be enough for the replication, way too small to ensure sufficient power. Tversky and Kahneman drew the following conclusion (1971, 105):

Apparently, most psychologists have an exaggerated belief in the likelihood of successfully replicating an obtained finding. (…) people have strong intuitions about random sampling; (…) these intuitions are wrong in fundamental respects; (…) these intuitions are shared by naive subjects and by trained scientists; and (…) they are applied with unfortunate consequences in the course of scientific inquiry.

That is, psychologists don’t seem to realize that except for very large samples any sample can differ substantially from the population it is drawn from, and that as a result two samples are thus unlikely to be very similar.

 To my knowledge, no similar survey has been conducted with scientists from other disciplines, but there is little reason to be optimistic about their better grasp of sampling error or about them taking it into account properly. Indeed, the consequence of the neglect of sampling error is found in many other sciences, as is shown in Section 3.3.

*3.3 Power Outage*

Beginning with psychology, as early as 1962, statistician Jacob Cohen was concerned with the power of experiments in psychology: The median power to detect a medium effect size in the 1960 volume of the *Journal of Abnormal and Social Psychology* was only .46. Many studies have confirmed this surprisingly low figure, and the situation has unfortunately not improved since the 1960s. The median power for a medium effect size was .37 for the articles published in the 1984 issue of this very same journal (Sedlmeier and Gigerenzer, 1989). The sample sizes in developmental psychology are typically only adequate to observe very large effects with reasonable power (Oakes, 2017). In social psychology, in no leading journal, including the *Journal of Personality and Social Psychology*, the most influential journal in this area of psychology, was the median power for a medium effect size larger than .5 for a sample of articles randomly selected from the 2006-2010 volumes; in personality psychology, things were slightly brighter, although not bright, with a median power around .6 (Fraley and Vazire, 2014).

Neuroscience does not fare better: The median power to detect the metaanalytically estimated effect size in 730 studies published in 2011 across several areas of neuroscience (neuroimaging, animal model studies, candidate gene association studies) is around 20%, with the exception of a handful of extremely highly powered studies in neurology (Button et al., 2013). A reanalysis of this data set has shown that power varies substantially across and also within neuroscientific disciplines (Nord et al., 2017), but power remains insufficient for most studies. An analysis of t-tests in nearly 4,000 papers from 18 leading journals in cognitive neuroscience (e.g., *Nature Neuroscience, Neuron, Brain, The Journal of Neuroscience, Cerebral Cortex, NeuroImage, Cortex*) and in psychology (e.g., *Psychological Science, Cognitive Science, Cognition*) over a four-year period (2011-2014) estimated a power to detect a medium effect equal to .44, with cognitive neuroscience being in worse shape than psychology (Szucs and Ioannidis, 2017). Similar figures are found in behavioral ecology (Jennions and Møller, 2003).

Even more depressing, medicine fares poorly. Clinical trials test the putative effects (improving health, preventing infection, toxic side effects, etc.) of biomedical interventions, and their success or failure can lead to the public release of a new treatment or, equally, to the decision not to release it. When the population effect size is taken to be equal to its metaanalytic estimate, less than 7% of 130,000 clinical trials from 1975 to 1914 have a power higher than .8 (Lamberink et al., 2018). Needless to say, it is immoral to expose participants to experimental drugs when little can be known from the studies. Relatedly, sample sizes remain extremely low for many clinical trials: More than half of the clinical trials preregistered on ClinicalTrials.gov included a sample smaller than 100 participants (Califf et al., 2012). Association biomedical studies, which examine the association of biological, environmental or cognitive parameters with neurological, psychiatric, and somatic diseases (e.g., autism, Alzheimer, etc.), are equally worrying: Using 660 metaanalytic estimates to compute power for a 4 year period (2008-2012), median power was very low for biological parameters—“17% for somatic diseases, 20% for psychiatric disorders and 20% for neurological diseases”—but satisfactory for cognitive and behavioral parameters related to ADHD, depression, and schizophrenia (Dumas-Mallet et al., 2017). Genetic studies had a very low median power (.08) and neuroimaging studies a better, but still unsatisfying one (.27).

Of course, the power of experiments depends, among other things, on the true population effect size: Even if experiments would have a low power if effect sizes were low, perhaps their actual power is large because the effects to be detected are themselves large. Indeed, Cohen (1962) explained the ignorance of power by speculating that psychologists believe that population effect sizes are large. And some have suggested they indeed are! Early analyses of power in neuroimaging argued that neuroimaging studies were sufficiently powered despite being small because they were detecting very large effects (e.g., Desmond and Glover, 2002). It is however unreasonable to expect effects to be large. Overviews of decades of studies in several sciences have shown that effect sizes are often small, and require very large sample sizes to be reliably detected. The mean effect size in 100 years of studies in psychology (combining more than 300 metaanalyses, 25,000 studies, and eight million participants!) appears to be equal d=.4 (Richard, Bond, and Stokes-Zoota, 2003). A third of the effects examined in this study are very small (less than d=.2) and only a quarter substantial (d>.6). Effect sizes are no larger in epidemiology: .2 for the more than 100,000 studies examined by Lamberink et al. 2018.

*3.4 Power Outage and the Fragility of Empirical Literatures*

Power matters in several crucial respects for the trustworthiness of science. Most obviously, a study with low power that tests a true hypothesis is likely to result in a false negative. Resources invested to conduct this study are thus likely to be wasted, which raises ethical concerns when these resources include human beings as well as public funding. More to the point of this article, low power underwrites the fragility of empirical literatures: It results in the frequency of false positives in published literatures, and it causes failed replications to be uninformative. I examine each point in turn.

The rate of false positives among significant results varies inversely with their average power, and a low power means, for many plausible rates of true hypotheses, a high proportion of false positives. This relation is captured by the false positive rate (e.g., Ioannidis, 2005; Button et al., 2013; Benjamin et al., 2018). The false positive rate is a function of the prior probability that the null hypothesis is true (the number of true null hypotheses under test), represented here by *φ*, the significance level (α), and the power (1-β):

1. $FPR=\frac{αφ}{αφ+(1-β)(1-φ)}$

As is clear from (1), as power decreases the proportion of false positives among significant results increases. The low replicability of some literatures is in part a reflection of this state of affairs since most empirical literatures are only constituted of significant findings (Fanelli, 2010). Figure 2 presents various false positive rates as a function of the frequency of true nulls and the power of experiments.



Figure 2: False Positive Rates

Second, replications are likely to be underpowered, first because the effect size of the original study is likely to be overestimated (small studies overestimate effect sizes), second because scientists do not take sampling error sufficiently into account. As a result, even if the hypothesis under test is correct, the resulting literature may contain apparently contradictory results from original studies and replications (e.g., Machery, Grau, and Pury, 2020): The original result is a true positive, but the failed replication is a false negative due to low power.

*3.5 The Permanence of Power Outage and the Neglect of Sampling Error*

Why do underpowered studies persist? Why have scientists disregarded methodologists’ repeated advice for so long? Why have official recommendations been ignored? The 2010 edition of the *APA* manual went out of its way to require power computation (American Psychological Association, 2010, 30; see also the guidelines from the Psychonomic Society’s Publications Committee and Ethics Committee in 2014): “When applying inferential statistics, take seriously the statistical power considerations associated with your test of hypotheses.” To no avail. Not only that, why haven’t scientists listened to mother nature’s feedback when she told them again and again there was something wrong in their repeatedly failed experiments?

Cohen expressed his puzzlement in his influential “Power Primer” article in 1992: “It is not at all clear why researchers continue to ignore power analysis” (Cohen, 1992, 155). Power analysis may have been too complicated to become part of the toolkit of practicing scientists; the notion of power may be part of a statistical practice that is so confused that it can only engender misuses (Sedlmeier and Gigerezer, 1989); or scientists may not have been fully aware of the importance of effect sizes, which are needed to compute power (Cohen, 1992). However, none of these candidate explanations successfully explains the decade-long permanence of low power in science. Scientists have a poor grasp of other statistical notions such as p-value or confidence intervals (e.g., Oakes, 1986; Belia, Fidler, Williams, and Cumming, 2005), but don’t neglect them. The notion of p-value belongs to the same statistical practice as the notion of power, but isn’t neglected. And while effect sizes were neglected for decades in many disciplines, this isn’t really the case anymore.

Another, recent explanation appeals to widespread, extensive p-hacking: Power doesn’t matter when you p-hack. Scientists are not troubled by the low power of their studies because they are able to obtain significant results systematically by p-hacking them (Nelson et al., 2018). In particular, when a study runs many tests, any of these tests can have low power, while it is likely that one of the tests will get a significant result: If there are only 20 participants per cell, the power of obtaining at least one effect in a 2x2 omnibus ANOVA is .93 for a medium effect size. Experimental projects that test many hypotheses, a form of data analytic flexibility, are thus likely to be successful and lead to publication. However, while p-hacking is real, it is only part of the story: There is no evidence that scientists p-hack extensively, in contrast to occasionally. And a moderate p-hacking still leaves it mysterious why scientists haven’t reformed their way, and taken power more seriously.

So, instead of the difficulty of the concept of power and instead of p-hacking, I propose that the neglect of sampling error largely explains the permanent low power observed in psychology and other sciences, and as a result the low replicability low power engenders. If scientists really underestimate sampling error, as they actually do (see the discussion of Kahneman and Tversky above), they would design insufficiently powered experiments. And indeed one can easily trace the causal link from a neglect of sampling error to the design of insufficiently powered experiments: When asked to determine how much power is desirable, psychologists respond on average .80, but when asked the size of the samples they typically collect and the effect size they expect, the resulting power is well below .8; when presented with a sample size and a population effect size, psychologists and neuroscientists typically overestimated by 50% to 100% the resulting power (Bakker et al., 2016).

But, one might respond, how is it possible that scientists haven’t yet learned that their experiments are insufficiently powered since such experiments are likely to yield negative results, and thus fail by their own lights? If they p-hacked extensively, such failures wouldn’t be an issue, but I have rejected the claim that scientists p-hack extensively. Instead of p-hacking, scientists are able to see past failed experiments because of their attitude toward negative results and the willingness to shelve studies in the file drawer. They blame everything but the power of the study when it fails. In particular, they believe that failed experiments reveal that aspects of the experimental procedure itself, such as the manipulation or the measurement, but not the sample size, were deficient or that the sample collected was in some respect or other defective: For instance, participants may have been unfortunately inattentive or uncooperative. And they can then shelve these failed experiments in the file drawer, having rationalized, although misidentified, the source of their failure.

To summarize, sampling error is not intuitive, and scientists do not seem to have a good grasp of it. The neglect of sampling error plausibly leads scientists in many disciplines to design insufficiently powered experiments, and this low power underwrites the fragility of empirical literatures, by producing a high false positive rate and by leading scientists to conduct uninformative replications.

As noted in the introduction, this conclusion is in line with the error-theoretic approach in philosophy of science, which emphasizes the role of error detection and correction in the development of trustworthy science. In particular, the neglect of sampling error discussed in this section falls under one of the types of errors identified by Allchin (2001): “sampling error,” a form of “observational error.” What is new in the present discussion is connecting this neglect to the current replication crisis and what it reveals about the untrustworthiness of some scientific literatures.

**4. The Neglect of Measurement Error**

A second aspect of the mistaken confidence in data is found in scientists’ neglect or poor grasp of random and systematic measurement error. In this section, I argue that the neglect of measurement error may contribute to the emergence of ghost literatures and could be another source of the instability of empirical literatures. “Random error” refers, roughly, to the fluctuations in measurement around the true value; random error is probabilistically distributed (often normally distributed) around this value. “Systematic error” refers, roughly, to the directional component of the measurement scores that is repeatable across measurements. “Neglect” is used as previously defined, and continues to have an epistemic and a doxastic source.

*4.1 The Frequency of Mismeasurement*

I will call “mismeasurement” the set of practices that embody a neglect of measurement error, and as a result undermine successful, i.e., reliable and valid, measurement in science.

Measurement in psychology is often treated with too little care. Psychologists multiply measures meant, arguably, to measure the same construct. Emotions are an important psychological construct that plays an important role in both lay and scientific psychological explanations of human behavior, but there is no accepted, validated instrument to ask participants to report their occurrent emotional states (in contrast to their disposition to experience some particular emotion) (Weldman, Steckler, and Tracy, 2016). Rather, the occurrence of any given emotion is assessed by asking several distinct questions, and the same question is often used to assess the occurrence of several emotions: For instance, joy, excitement, schadenfreude, and happiness are all measured by asking a participant whether she is happy. Lest you think the measurement of emotions is an exception, consider the measure of explicit racism, the kind of racism that people are aware of and able to avow: A recent review compared nearly 50 different ways of measuring explicit racism (Axt, 2018). “Harder” areas of the behavioral sciences do not fare much better: There are more than 280 scales and tools to measure depression.

 The existence of many different measures would not be an issue if their convergent and discriminant validities were known, but this is not the case for many measures in psychology. As a result, their construct validity is at best questionable. Roughly, two measures have discriminant validity to the extent that they are meant to measure different constructs and their scores are not highly correlated. For instance, a measure of explicit bias and a measure of implicit bias, which are meant to measure different things, would have discriminant validity to the extent that their scores were not correlated. Roughly, two measures have convergent validity to the extent that they are meant to measure the same construct and that their scores are correlated. For instance, two measures of depression have high convergent validities if their scores are highly correlated. When a measure has low convergent validity with other measures meant to measure the same construct and low discriminant validity with measures meant to measure other constructs, then it does not measure well what it is supposed to measure: It has low construct validity.

 Despite limited knowledge of psychological measures’ convergent and discriminant validities, scientists often use measures interchangeably when they are meant to measure the same construct and treat measures meant to measure different things as if they are actually measuring different things. The Jangle-Jingle fallacy could explain this lackadaisical attitude (Thorndike, 1904; Weldman et al., 2016). Someone commits the Jingle fallacy when she takes two things to be the same because they have the same name and the Jangle fallacy when she takes two things to be different because they have a different name. Calling different measures “measures of depression” prime psychologists, psychiatrists, and neuroscientists to assume that they all measure the same thing—an instance of the Jingle Fallacy. Calling a measure a “measure of implicit racism” and another a “measure of explicit racism” primes psychologists to assume that these two measures measure different things—an instance of the Jangle Fallacy.

 In addition to the proliferation of measures, another form of mismeasurement is the use of ad hoc measures, i.e., measures that are invented for a given research project and whose psychometric properties are not evaluated. One third of the articles in the 2014 volume of the *Journal of Personality and Social Psychology* contain ad hoc measures; nearly half of the measures are used without citation, suggesting that authors do not expect their readers to be concerned with the reliability and validity of the measures they use (Flake, Pek, and Hehman, 2017). The use of ad hoc measures reveals that psychologists do not think the introduction of a new measure requires extensive research to assess its psychometric properties. Another sign of the lack of attention to the validation of measures is the fact that many measures are modified, but not revalidated: 19% of the measures used in the 2014 volume of the *Journal of Personality and Social Psychology* and drawn from previous work were modified without revalidation (Flake et al., 2017).

 Am I painting a dire picture of psychological measurement by focusing on measures that are not particularly well established? Some measures are indeed on a much better footing than many of the measures used in psychology, but even those measures are not beyond concern (Hussey and Hugues, 2020). Many of these measures have good internal consistency and test-retest reliability: Their components agree with one another and when measurement is repeated the obtained scores are correlated. Other important characteristics of good measures are however less satisfying. The factor structure of many scales and their measurement invariance are often unsatisfying. Factor analysis is used to identify how the components of a scale cluster together. Their clustering is supposed to be in line with the dimensions of what they are meant to measure. For instance, according to the OCEAN theory of personality, personality has five main dimensions: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. The factor structure of the questionnaire is good to the extent that its items cluster around five factors in the way theoretically expected. The factor structure of personality tests is apparently mixed (Hussey and Huges, 2020). Measurement invariance holds when the psychometrics properties of scales remain the same across demographic groups. The measurement invariance of personality tests has recently been under extensive discussion, and the matter is not settled.

*4.2 The Neglect of Measurement Error and the Replication Crisis*

The neglect of measurement error has contributed to the replication crisis in various ways. I focus on two issues here (for a third point, see Loken and Gelman, 2017).

 Poor validity contributes to the second feature of the replication crisis: ghost literatures. In the usual typology of replications, conceptual replications modify the operationalization of the manipulations (independent variables) or measurements (dependent variables). Let’s focus on measurement. When a measure has low construct validity, measurement variance results less from the variance of the construct it is meant to measure than from the variance of other constructs. When the construct validity of a measure is unknown, measurement variance could largely result from the variance of constructs the measure isn’t meant to measure. Thus, it is unknown whether two measures that are supposed to measure the same construct, as happens in conceptual replications, actually measure the same construct when their validities are unknown. They might measure different things although this possibility is hidden by the Jingle fallacy. This situation undermines the usefulness of conceptual replications. When a conceptual replication fails and when the construct validities are unclear, it may be because the measure used in the replication does not measure the same thing. When a conceptual replication succeeds, it may be that the first study was a false positive and the second study, measuring another construct, a true positive (or vice-versa). In fact, the latter situation contributes to explaining the emergence of ghost literatures. Ghost literatures are based on conceptual replications, as was noted in Section 2.[[1]](#footnote-1) This accumulation could result from questionable research practices, but it can also result from a mix of true and false positives due to the use of measures with poor construct validities.

 Second, poor measurement invariance means that empirical results are sample-dependent. A measure is invariant to the extent that its psychometric properties remain the same across groups. At least three psychometric properties are relevant: a measure’s reliability, its factor structure (when it is made of many components), and its factor loadings. When these psychometric properties vary across demographic groups, e.g., gender, SES, generations, or cultures, then empirical results are hostage of the demographic composition of the original and replication samples: Even if the original results are true positives, replications may fail when the replication sample differs demographically from the original study’s sample. If reliability decreases with the composition of the replication’s sample, it is harder to detect a genuine effect. If the factor structure or loading changes, then an intervention that strongly influences one of the factors of the measure may not have any measurable impact in a replication. Empirical results are thus bound to be unstable, creating uncertainty about the empirical record.

 To summarize, the neglect of measurement error leads scientists to ignore the poor psychometric properties of their measurements, including poor reliability, poor construct validity, and poor measurement invariance. Poor construct validity may contribute to the emergence of ghost literatures, while poor measurement invariance to the fragility of empirical literatures.

**5. Conclusion**

Scientists and science consumers should learn to distrust data, not because they are often made up (although they sometimes are), but because of how much information they really carry. Data are constantly infected by uncertainties, due to sampling or measurement error, and it is easy to overlook them. Neglecting sampling error results in fragile empirical literatures, due to a high proportion of false positives among significant results, overestimated effect sizes, and uninformative replications; neglecting measurement errors is a further source of the instability of empirical literatures because measurement may not be invariant and a possible source for the emergence of ghost literatures.

 Policy makers and the general public more generally consume scientific reports and, ideally, use them to inform their decisions. In light of the neglect of measurement and sampling errors discussed in this article, one might wonder whether they should adopt a more skeptical attitude toward science. Although addressing this question with the required amount of care is a matter for a distinct article, the onus of carefully assessing the uncertainties of empirical results should fall squarely on the shoulders of the scientific community. At an abstract level, the norm of organized skepticism in science should be reinforced: While scientists always try to balance trust and skepticism, the latter norm should be reinforced. More concretely, scientific and statistical education should highlight the uncertainty involved in producing scientific knowledge, from measurement, to sampling, to statistical modeling, and to theoretical conclusions. Even more concretely, the list of potential errors that scientists are attuned to should be broadened, to include greater emphasis on sampling and measurement errors.

Policy makers and the general public are even less likely than scientists to be appropriately sensitive to the limitations of the data collected by scientists, miscalibrating their skepticism: either by being too skeptical or by being too credulous. Assessing uncertainty shouldn't fall on their shoulders. On the other hand, scientists are incentivized to exaggerate the significance of their research in their interactions with policy makers, who fund them, and the lay public. Barring a radical, improbable change in these incentives, policy makers and the lay public should be educated about the frailties of on-going science.

One might worry that this message could reinforce the already strong skeptical undercurrent toward science in the general public. While this is a serious concern, developing a more guarded attitude toward science in the lay public may have benefits: By avoiding a credulous attitude toward science, the lay public is less likely to be disappointed when confronted with the challenges scientists cannot but encounter.

**Conflict of Interest**

The author declares that he has no conflict of interest.

**References**

Allchin, D. (2001). Error types. *Perspectives on science*, *9*(1), 38-58.

Axt, J. R. (2018). The best way to measure explicit racial attitudes is to ask about them. *Social Psychological and Personality Science, 9*(8), 896-906.

Baker, M., & Dolgin, E. (2017). Cancer reproducibility project releases first results. *Nature*, *541*,269-270.

Begley, C.G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*, 531‐533.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological methods, 10*(4), 389-396.

Benjamin, D. J., et al. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10.

Button, K.S., et al. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Review Neuroscience, 14*, 365376.

Califf, R.M., et al. (2012). Characteristics of clinical trials registered in ClinicalTrials. gov, 2007-2010. *Jama*, *307*, 1838-1847.

Chang, A., & Li, P. (2015). Is economics research replicable? Sixty published papers from thirteen journals say “usually not”. *Available at SSRN 2669564*.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.

Desmond, J.E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. *Journal of neuroscience methods*, *118*(2), 115-128.

Dumas-Mallet, E., et al. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society open science*, *4*(2), 160254.

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PloS one, 5*(4), e10068.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11), 2628-2631.

Flake, J.K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Fraley, R.C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one, 9(10)*, e109019.

Hagger, M. S., et al. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology, 14*(11), e2000995.

Hon, G. (1989). Towards a typology of experimental errors: An epistemological view. *Studies in History and Philosophy of Science Part A*, *20*(4), 469-504.

Hussey, I., & Hughes, S. (2020). Hidden invalidity among fifteen commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science,* *3*(2), 166–184.

Jennions, M.D., & Møller, A.P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*, *14*(3), 438-445.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Med, 2(8)*, e124.

Lamberink, H.J., et al. (2018). Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. *Journal of clinical epidemiology*, *102*, 123-128.

Lemoine, N. P., et al. (2016). Underappreciated problems of low replication in ecological field studies. *Ecology*, *97*(10), 2554-2561.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

Machery, E. (2015). Cognitive penetrability: A no-progress report. In J. Zeimbekis and A. Raftapoulos (Eds.), *The cognitive penetrability of perception* (pp. 59-74). Oxford: Oxford University Press.

Machery, E. (2020). What is a replication?. *Philosophy of Science*, *87*(4), 545-567.

Machery, E. Grau, and Pury, C. (2020). Love and power: Grau and Pury (2014) as a case study of the challenges in x-phi replication. *Review of Philosophy and Psychology, 11*, 995–1011.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Nelson, L.D., Simmons, J., and Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology*, *69*, 511-534.

Nord, C.L., Valton, V., Wood, J., & Roiser, J.P. (2017). Power-up: a reanalysis of “power failure” in neuroscience using mixture modeling. *Journal of Neuroscience*, *37*(34), 8051-8061.

Nosek, B.A., Spies, J.R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester, England: Wiley.

Oakes, L.M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22*(4), 436-469.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*, aac4716. doi: 10.1126/science.aac4716.

Richard, F.D., Bond Jr, C.F., & Stokes-Zoota, J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331-363.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–316.

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological science*, *28*, 687-693.

Szucs, D., & Ioannidis, J.P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS biology*, *15*(3).

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science, 22*(11), 1359-1366.

Thorndike, E.L. (1904). *An Introduction to the Theory of Mental and Social Measurements*. New York: Teachers College, Columbia University

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105-110.

Weidman, A.C., Steckler, C.M., & Tracy, J.L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267-295.

Windish, D. M., Huot, S. J. &, Green, M. L. (2007). Medicine residents' understanding of the biostatistics and results in the medical literature. *JAMA, 298*, 1010-1022.

1. To clarify, ghost literatures are not defined by this particular etiology. Presumably, many factors contribute to the existence of ghost literatures, and different ghost literatures may have different etiologies. [↑](#footnote-ref-1)