Philos Stud (2021) 178:133–145
https://doi.org/10.1007/s11098-020-01424-z

Inflated effect sizes and underpowered tests: how the severity measure of evidence is affected by the winner's curse

Guillaume Rochefort-Maranda[1]

Published online: 29 January 2020

Abstract My aim in this paper is to show how the problem of inflated effect sizes (the Winner's Curse) corrupts the severity measure of evidence. This has never been done. In fact, the Winner's Curse is barely mentioned in the philosophical literature. Since the severity score is the predominant measure of evidence for frequentist tests in the philosophical literature, it is important to underscore its flaws. It is also crucial to bring the philosophical literature up to speed with the limits of classical testing. The Winner's Curse is one of them. The problem is that when a significant result is obtained by using an underpowered test, the severity score becomes particularly high for large discrepancies from the null-hypothesis. This means that such discrepancies are very well supported by the evidence according to that measure. However, it is now well documented that significant tests with low power display inflated effect sizes. They systematically show departures from the null hypothesis H0 that are much greater than they really are. From an epistemological point of view this means that a significant result produced by an underpowered test does not provide evidence for large discrepancies from H0. Therefore, the severity score is an inadequate measure of evidence. Given that we are now aware of the phenomenon of inflated effect sizes, it would be irresponsible to rely on the severity score to measure the strength of the evidence against the null. Instead, one must take appropriate measures to try and avoid using underpowered tests by setting a threshold for the sample size or by replicating the results of the experiment.

Keywords Confirmation · Evidence · Statistics · Severity · Error statistics · Winner's curse

✉ Guillaume Rochefort-Maranda
guillaumemaranda@hotmail.com

[1] Laval University, Québec City, Canada

1 Introduction

In philosophy of statistics, Deborah Mayo and Aris Spanos have championed the following epistemic principle, which applies to frequentist tests:

> **Severity Principle (full)**. Data x_0 (produced by process G) provides good evidence for hypothesis H (just) to the extent that test T severely passes H with x_0. (Mayo and Spanos 2011, p.162).

They have also devised a severity score that is meant to measure the strength of the evidence by quantifying the degree of severity with which H passes the test T (Mayo and Spanos 2006, 2011; Spanos 2013). That score is a real number defined on the interval [0,1].

My aim in this paper is to show how the problem of inflated effect sizes (the Winner's Curse) corrupts the severity measure of evidence. This has never been done. In fact, the Winner's Curse is barely mentioned in the philosophical literature.[1] Since the severity score is the predominant measure of evidence for frequentist tests in the philosophical literature, it is important to underscore its flaws. It is also crucial to bring the philosophical literature up to speed with the limits of classical testing. The Winner's Curse is one of them.

The problem is that when a significant result is obtained by using an underpowered test, the severity score becomes particularly high for large discrepancies from the null-hypothesis. This means that such discrepancies are very well supported by the evidence according to that measure.

However, it is now well documented that significant tests with low power display inflated effect sizes when H1 is true. They systematically show departures from the null hypothesis H0 that are much greater than they really are: "theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated"(Ioannidis 2008, p.640) This is problematic in research contexts where the true discrepancies from H0 are particularly small and where the sample sizes are also small. See (Button et al. 2013; Ioannidis 2008; Gelman and Carlin 2014) for examples.

From an epistemological point of view this means that a significant result produced by an underpowered test does not provide evidence for large discrepancies from H0. Therefore, the severity score is an inadequate measure of evidence.

Given that we are now aware of the phenomenon of inflated effect sizes, it would be irresponsible to rely on the severity score to measure the strength of the evidence against the null. Instead, one must take appropriate measures to try and avoid using underpowered tests by setting a threshold for the sample size or by replicating the results of the experiment.

[1] I have searched with the key words "winner's curse" on the springer journal , the Oxford academic journal and on the university of Chicago press journal websites, with a filter on philosophy journals. I have also searched for the same key words on the Philpaper website. I only found one relevant reference and it is not in connection with the severity score: (Vieland and Chang 2018).

Unfortunately, the idea of increasing the power of a test in order to strengthen the evidence against the null is incompatible with Spanos and Mayo's claims to the effect that there is a common fallacies "wherein an α level rejection is taken as more evidence against the null, the higher the power of the test" (Mayo and Spanos 2006, p.344).

This paper contains two main sections. In the first section, I explain the problem of inflated effect sizes generated by underpowered tests with more details. I also provide an example using a Student's t-Test. In the final section, I explain why the severity score is an inadequate measure of evidence.

2 The argument and the methodology

The main argument that I put forward in this paper is very simple.

- An observed test statistic will display a misleading large departure (large effect size) from H0 when an underpowered test is significant and H1 is true.
- The severity score justifies larger discrepancies from the null when the observed effect size is larger and the test is significant.
- Therefore, the severity score is a measure that will be systematically wrong when evaluating the result of an underpowered test when H1 is true and the test is significant.

The premises of this argument are now established facts. The first premise more particularly is a well-known phenomenon:

> when an underpowered study discovers a true effect, it is likely that the estimate of the magnitude of that effect provided by that study will be exaggerated. This effect inflation is often referred to as the Winner's Curse (Button et al. 2013, p.366).

and it affects real scientific practice (e.g. neuroscience). It is not merely a theoretical problem:

> Our results indicate that the average statistical power of studies in the field of neuroscience is probably no more than between 8% and 31%, on the basis of evidence from diverse subfields within neuro-science. If the low average power we observed across these studies is typical of the neuroscience literature as a whole, this has profound implications for the field. A major implication is that the likelihood that any nominally significant finding actually reflects a true effect is small. (Button et al. 2013, p.371).

The originality of this paper is to put the first and second premise together in order to dismiss the severity score as an adequate measure of evidence. The purpose of this paper is not to prove the first premise nor to show that it is a real problem. As one can see, this has already been done. See (Ioannidis 2008). I will however illustrate the problem with a detailed example. The example that I provide is helpful because it shows the pervasive effect of the Winner's curse on the severity score given that the real discrepancy from H0 is known. We would not be able to do that with a real case-study because we usually do not know the real power of a test.

2.1 Inflated effect sizes generated by underpowered tests

The fact is that the lower the power of a test, the more the distributions of the test statistic under H0 and under H1 are similar. Consequently, the more extreme a test statistic must be under H1 (assuming that we know the true power of the test) in order to trigger a significant result. In fact, assuming that the power of a test is as low as 0.08 (see previous quote), a significant result provided by a low powered test will necessarily[2] display a departure from what we expect under both H0 and H1 (given the true power of the test). Consequently, in a context where we do not know the true power of a test, we will have the illusion that the true difference from H0 is greater than it really is if we rely on the value of the test statistics to justify the magnitude of the difference.

There are two necessary conditions to observe this phenomenon: significance and low power.

> Inflation is expected when, to claim success (discovery), an association has to pass a certain threshold of statistical significance, and the study that leads to the discovery has suboptimal power to make the discovery at the requested threshold of statistical significance. Both conditions are necessary to inflate effect sizes.
> (Ioannidis 2008, p.641).

This problem is fairly easy to illustrate. Imagine that a statistician S has obtained two different samples of 10 independent and identically distributed observations: $(X_1, X_2, \ldots, X_{10})$ and $(Y_1, Y_2, \ldots, Y_{10})$. Their respective distributions are defined as follows:

(1) $\boldsymbol{X_i} \sim \mathcal{N}(\mu_1 = 1.01, \sigma_1^2 = 36)$
(2) $\boldsymbol{Y_j} \sim \mathcal{N}(\mu_2 = 1, \sigma_2^2 = 36)$

where μ represents the mean of a normal distribution and σ^2 its variance.

S only knows two things about the parameters of the two normal distributions:

(1) $\mu_1 > \mu_2$ or $\mu_1 = \mu_2$
(2) $\sigma_1 = \sigma_2$

[2] First of all, a test statistic always displays a departure from what we expect under H0 when the test is significant with a small α (0.05). This is because the probability of a statistic reaching the critical region is as small as α. Now, let us imagine that the power is low (0.08) and also claim that a test statistic does not display a departure from that we expect under H1 when the test is significant. This means that we would expect the test statistic to reach the critical region under H1. But we just said that we do not expect this to be the case. The probability is 0.08. Therefore, it is impossible for a significant result to fail to display a departure from what we expect under both H0 and H1.

She does not know their exact value. Consequently, in order to make an inference about the difference between μ_1 and μ_2, S uses a one-tailed Student's t-Test where H1: $\mu_1 > \mu_2$ and H0: $\mu_1 = \mu_2$. The variances are estimated with the samples.

The statistic used for such a test is defined as follows:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

where

$$S_p = \sqrt{\frac{9S_1^2 + 9S_2^2}{18}},$$

$$S_1^2 = \sum_{i=1}^{10} \frac{((x_i) - \bar{X})^2}{9},$$

$$\bar{X} = \sum_{i=1}^{10} \frac{x_i}{10},$$

$$S_2^2 = \sum_{i=1}^{10} \frac{((y_i) - \bar{Y})^2}{9},$$

and

$$\bar{Y} = \sum_{i=1}^{10} \frac{y_i}{10}.$$

It is called a Student's t-Test because the statistic t follows a Student distribution (with 18 degrees of freedom in this case).

For a significance level α of 0.05, S will reject H0 (accept H1) if she finds a test statistic t_{obs} such that the probability of obtaining a result at least as distant (on the positive axis) from 0 as t_{obs} is smaller than or equal to 0.05 under H0. If not, then she will fail to reject H0.

The probability that will determine the rejection (or non-rejection) of H0 is called "the p value". In this particular case, α is the probability of rejecting H0 when H0 is true. It is also called "the probability of making a Type-I error". The probability of rejecting H0 when H1 is true is called "the power of the test" (π) and the probability of not rejecting H0 when H1 is true is "the probability of making a Type-II error" ($\beta = 1 - \pi$). In this case, the power of the test is very low given the small difference between the populations, the high variances and the small sample size.

In short, S expects the statistic t to be close to 0 under H0 because there should not be any difference between the two distributions. If the test statistics is much bigger than 0, then she will reject H0 and accept H1 because that would be too improbable under H0. If it is relatively close to 0, then she will not reject H0 because that is not too improbable under H0.

After S proceeds with the t-test, she finds a difference of 4.250; a test statistic $t_{obs} = 1.914$; and a p value = 0.036 (See "Appendix" to reproduce the results). Therefore, S rejects H0 (p value < 0.05). The test is significant.

In fact, the result is quite remarkable. S has observed a difference between the two means of 4.250 when the true difference is only 0.01. This is because we have a significant result with an underpowered test such that the effect size is incredibly bigger than reality (450 times greater). S would thus be wrong to believe that there is such a substantial difference from H0. But S would feel warranted to reach a similarly bad conclusion with the severity score.

2.2 The severity score

Suppose that S would like to use the severity score for $\mu_1 - \mu_2 > 0.1$. That score consists in a postdata evaluation of a test: "Severity constitutes a postdata evaluation of the Neyman–Pearson accept/reject results with a view to establish the smallest/largest discrepancy γ from H0 warranted by data x_0" (Spanos 2013, p.86). It is always evaluated for a given discrepancy of interest.

It can be used in order to assess if the evidence for a given discrepancy is good by looking at whether or not the severity score is above a certain threshold: "As we just saw, the statistically significant result, x=0.4, is good evidence for $\mu > .2$ (the severity was .841), but poor evidence for the discrepancy $\mu > .5$ (the severity was .3)" (Mayo and Spanos 2011, p.173). For this reason it can be seen as a measure of the strength of the evidence.

The severity score can also be seen as a measure of warrant for a given discrepancy from the null:

> "Now, let us consider the two other statistically significant outcomes, retaining this same inference of interest. When x=.6, we have SEV $(\mu > .2) = .977$, since x=.6 is 2 standard deviation in excess of the $\mu = .2$. When x=1, SEV $(\mu > .2) = .999$, since x=1 is 4 standard deviation in excess of $\mu = .2$. So inferring the discrepancy $\mu > .2$ is increasingly warranted, for increasingly significant observed values". (Mayo and Spanos 2011, p.170).

Since S is interested in $\mu_1 - \mu_2 > 0.1$, she decides to use the severity score to see how strong is her evidence for that claim or how warranted would be the inference for that claim.

She computes that score as follows:

$$t_s = \frac{(4.250) - (0.1)}{S_p \times \sqrt{\frac{1}{10} + \frac{1}{10}}}$$

$$SEV(\mu_1 - \mu_2 > 0.1) = \boldsymbol{F}(\boldsymbol{t_s}) = \mathbb{P}(T_s \leq \boldsymbol{t_s}) = 0.961$$

where $\boldsymbol{F}(\boldsymbol{t_s})$ is the cumulative distribution function of a Student's distribution with 18 degrees of freedom evaluated at point t_s.

In English, this means that S has computed the probability of obtaining a less extreme result under the assumption that $\mu_1 - \mu_2 = 0.1$. This is the meaning of the severity score in this context. See (Mayo and Spanos 2011, p.169) for even more details on how to compute such a severity score.

If the severity score is high, then we can infer that the data provides good evidence for $\mu_1 - \mu_2 > 0.1$ and believe that inferring that there is such a discrepancy from the null is warranted. This is the case here and it should not come as a surprise given that S has observed such an inflated effect size. Notice that the severity score will be higher the greater the observed size effect (just look at the numerator of the fraction that generates t_s).

In a nutshell, S has found a significant result (p value $= 0.036$). She thus rejects H0 and finds a high severity score for the claim $\mu_1 - \mu_2 > 0.1$ (severity score=0.961). Hence, S believes that she has good evidence for such a difference that is at least ten times larger than the true difference.

However, S would be epistemically irresponsible to trust the severity score given what is now known about the problem of effect sizes and underpowered tests. If the severity score is high for $\mu_1 - \mu_2 > 0.1$, it is because the observed effect size is very big. Inflated effect sizes corrupt the severity measure of evidence.

In fact, S would not even be justified to believe that there is some discrepancy from the null because she would most likely not be able to replicate her results. As Gelman, Carlin, Ioannidis, Stanley and Doucouliagos put it:

> The problem, though, is that if sample size is too small, in relation to the true effect size, then what appears to be a win (statistical significance) may really be a loss (in the form of a claim that does not replicate) (Gelman and Carlin 2014, p.642).
>
> When power is low, reported statistically significant findings are quite likely to be artefacts from chance and bias (Ioannidis et al. 2017, p.f240).

In the example presented above, even if S were to repeat her experiment 100,000 times, she would not be able to obtain enough evidence to reject H0. To see this, 100,000 p values associated with 100,000 replications of the experiment are represented in Fig. 1 (See "Appendix" to reproduce the results).

Given that a p value follows a uniform distribution under H0 but not under H1, S could perform a Kolmogorov–Smirnov test for the uniformity of the p values. Doing

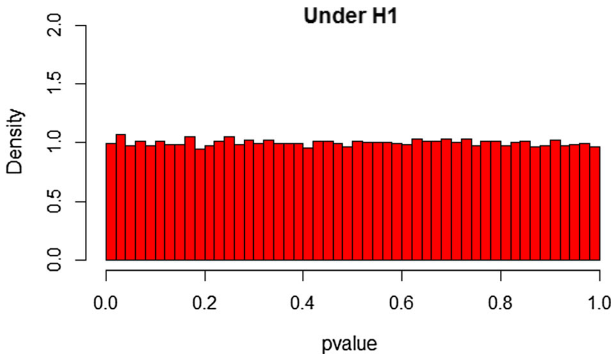


Fig. 1 Histogram estimation of the density of the p values, under the assumption that H1 is true, made with 100,000 simulations

so, she would obtain a test statistic of 0.002 and a p value of 0.958 (See “Appendix” to reproduce the results). This means that S would not be able to reject the hypothesis stating that those p values follow a uniform distribution. This also means that she would not be able to reject the hypothesis stating that the two means are equal.

In sum, the severity score is an inadequate measure of evidence and should be rejected. That score is sensitive to the inflated effect sizes provided by underpowered tests. In order to assess the strength of the evidence, one must make sure that a departure from the null is not an artefact of an underpowered test. The severity score is useless for that purpose.

One could hang on to that measure and claim that it is appropriate in contexts where the power is not too low. This is a possibility. But it means that the severity score does not capture the notion of “warrant” or “good evidence”. It is incomplete at best.

3 The more power the better

Naturally, in light of what has just been said, one must try to make sure that a test is powerful in order to generate good evidence against the null. “If the discovery studies were fully powered, inflation would not be an issue”(Ioannidis 2008, p.641). The more power the better.

Unfortunately, that solution is incompatible with the idea that it is fallacious to claim that a significance test provides more evidence against the null, the higher the power of the test (Mayo and Spanos 2006, p.344). One of the main lessons taken from the study of the Winner’s curse is that underpowered tests must be replaced with more powerful ones if we are to take their rejections of the null hypothesis seriously. In other words, those tests need to have more power in order to provide more evidence against the null. Underpowered tests are not replicable and do not provide evidence against H0.

This does not sit very well with the severity score. As it is shown in (Mayo and Spanos 2006, p.344), the severity score increases when the power decreases for a given test statistic. In fact, proponents of the severity score do not believe that more powerful tests can provide better evidence against the null simply because we can detect minute differences from the null if our tests are powerful enough. Indeed, it is often said that we can always reject H0 with enough observations. Hence, it would be wrong to conclude that there is an interesting discrepancy from H0 simply because we reject H0 with a powerful test.

But showing that a powerful test has only warranted the existence of a small discrepancy from H0 does not mean that we have little evidence against H0 and that H1 is not well supported by the evidence. The existence of a small difference from H0, if well justified, is enough evidence against H0. By analogy, a proof that a bone is sprained is not worse evidence against the hypothesis that there is no bone damage than a proof that a bone is broken.

There is a clear distinction between (1) claiming that a significant test provides justification for an scientifically interesting difference from H0 and (2) claiming that it provides justification for a difference of λ from H0. A small difference from H0 can be extremely well-justified. What inflated effect sizes show is that if we want to justify or warrant the existence of a difference λ (whatever it may be), then we need a significant result obtained with a powerful test.

4 Conclusion

In a nutshell, the severity score is an inadequate measure of evidence (or warrant) and should be rejected or considered incomplete at best. It is sensitive to the inflated effect sizes provided by underpowered significant tests when H1 is true. The point is that inflated effect sizes also inflate severity scores. Therefore, the severity score misleadingly warrants discrepancies that are much larger than the truth when the power is low. This has not yet been pointed out in the philosophical literature.

I have illustrated this with an example. In order to assess the strength of the evidence, one must make sure that a departure from the null is not an artefact of an underpowered test. One can do so by taking reasonable precautions against low powered tests, such as trying to replicate the results of a test.

Like it was mentioned in the introduction, the problem of inflated effect sizes provided by significant and underpowered tests is not merely a theoretical problem. The interested reader can consult (Gelman and Carlin 2014) who mention two specific examples taken from published work. This makes it all the more important to underscore the inadequacies of the severity score as a measure of evidence.

In sum, I have shown that the following quotes also applies to philosophy of science:

> it is not sufficiently well understood that "significant" findings from studies that are underpowered (with respect to the true effect size) are likely to produce wrong answers (Gelman and Carlin 2014, p.649).

Philosophers have overlooked the problem of inflated effect sizes. The Winner's curse is crippling the severity measure.

Appendix

The test with 10 observations per group

```
set.seed(31)
x<-rnorm(10, 1.01, 6)
y<-rnorm(10, 1, 6)
grp<-c(rep(1, 10), rep(2, 10))
z<-c(x, y)
dat<-as.data.frame(cbind(z, grp))
test<-t.test(z~grp, data=dat, var.equal=T, alternative = "greater")
test

##
##  Two Sample t-test
##
## data:  z by grp
## t = 1.914, df = 18, p-value = 0.03583
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.3994825       Inf
## sample estimates:
## mean in group 1 mean in group 2
##        2.688654       -1.560957
```

We find the severity score for a difference strictly larger than (0.1).

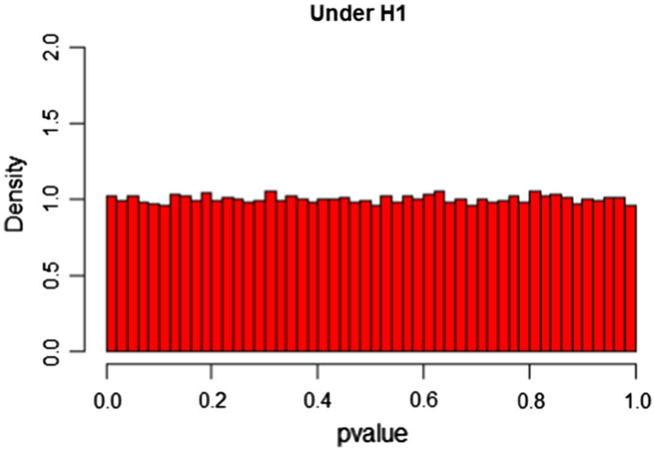
```
set.seed(31)
x<-rnorm(10, 1.01, 6)
y<-rnorm(10, 1, 6)
s1<-sum((x-mean(x))^2)/9
s2<-sum((y-mean(y))^2)/9
sp<-sqrt((9*s1+9*s2)/18)
a<-sqrt((1/10)+(1/10))
t<-((2.688654 + 1.560957)-(0.1))/(sp*a)
sev<-pt(t, df=18, lower.tail = T, log.p = FALSE)

sev

## [1] 0.9610043
```

We find the distributions of the p-value under the assumption that H1 is true for the test with 10 observations per group

```
 pvh1<-rep(NA, 100000)
for (i in 1:100000){
  x<-rnorm(10, 1.01, 6)
  y<-rnorm(10, 1, 6)
  grp<-c(rep(1, 10), rep(2, 10))
  z<-c(x, y)
  dat<-as.data.frame(cbind(z, grp))
  test<-t.test(z~grp, data=dat, var.equal=T, alternative = "greater")
  pvh1[i]<-test$p.value
#  print(i)
}
pvalue<-pvh1
hist(pvalue, freq=F, 50, ylim=c(0, 2), col=2, main="Under H1")
```



We perform a Kolmogorov-Smirnov test for the uniformity of the p-values under H1.

```
ks.test(pvh1, "punif")

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  pvh1
## D = 0.0016108, p-value = 0.9576
## alternative hypothesis: two-sided
```

Extra simulations

We perform a low powered test with a difference of 0.4 and compute the mean severity for a discrepancy of 0.4.

```
library(severity)
set.seed(7356581)
X = rnorm(1000, 0.4, 1)

sev1 = rep(0, 3000)

for(i in 1:3000){
  samp = sample(X, 25)
  result = t.test(samp, mu = 0)
  if(result$p.value < 0.05){
    sev1[i] <- severity(mu0 = 0, xbar = mean(samp), sigma = 1, n = 25, alpha
= 0.05)$severity_rejectH0[19]
  }else{sev1[i] <- NA}
  print(i)
}
mean1<-mean(na.omit(sev1))
```

We then perform a low powered test with a difference of 0.1 and compute the mean severity for a discrepancy of 0.4. We see that the mean severity score is larger than the previous one. This means that we can better justify a discrepancy that is 4X larger than the truth (0.1) than a true discrepancy of 0.4 when the power is low.

```
set.seed(735653281)
X = rnorm(1000, 0.1, 1)

sev = rep(0, 3000)

for(i in 1:3000){
  samp = sample(X, 20)
  result = t.test(samp, mu = 0)
  result
  if(result$p.value < 0.05){
    sev[i] <- severity(mu0 = 0, xbar = mean(samp), sigma = 1, n = 20, alpha = 
0.05)$severity_rejectH0[17]
  }else{sev[i] <- NA}
  print(i)
}
mean2<-mean(na.omit(sev))
mean1 = 0.7758278

mean2 = 0.8450679
```

References

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, J. E. S., et al. (2013). Power failure: Whysmall sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal, 127*(605), F236–F265.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357.

Mayo, D. G., & Spanos, A. (2011). Error statistics. *Philosophy of Statistics*, *7*, 152–198.

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science*, *80*(1), 73–93.

Vieland, V. J., & Chang, H. (2018). No evidence amalgamation without evidence measurement. *Synthese, 196*(8), 3139–3161.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.