Reframing the Reproducibility Crisis: Using an Error-Statistical Account to Inform the Interpretation of Replication Results in Psychological Research.

Caitlin Grace Parker

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Arts
In
Philosophy

Deborah Mayo
Benjamin Jantzen
Lydia Patton

May 6th, 2015
Blacksburg, VA

Keywords: statistics, replicability, reproducibility, psychology, scientific inference

Reframing the Reproducibility Crisis: Using an Error-Statistical Account to Inform the Interpretation of Replication Results in Psychological Research.

Caitlin G Parker

ABSTRACT

Experimental psychology is said to be having a reproducibility crisis, marked by a low rate of successful replication. Researchers attempting to respond to the problem lack a framework for consistently interpreting the results of statistical tests, as well as standards for judging the outcomes of replication studies. In this paper I introduce an error-statistical framework for addressing these issues. I demonstrate how the severity requirement (and the associated severity construal of test results) can be used to avoid fallacious inferences that are complicit in the perpetuation of unreliable results. Researchers, I argue, must probe for error beyond the statistical level if they want to support substantive hypotheses. I then suggest how severity reasoning can be used to address standing questions about the interpretation of replication results.

ACKNOWLEDGMENTS

I would like to take this opportunity to recognize several individuals who played an important role in the completion of this project.

I thank my advisor, Deborah Mayo, for her guidance, expertise, and unbelievable support. I am immensely grateful to her for her knowledgeable and detailed feedback, for her careful explanations, and for her encouraging words during times of hardship. I am also indebted to Benjamin Jantzen for his general insight and excellent constructive criticism, as well as for inspiring me to pursue philosophy in the first place. Similarly, I thank Lydia Patton for her help clarifying ideas and assistance navigating such a complicated project.

I would also like to recognize my colleagues for their contributions. I am particularly grateful to fellow graduate students John Waters, Stacey Kohls, and Shannon Abelson, along with program alumni Dan Linford and Joanna Roye, for their insightful comments and their incredible moral support.

Finally, none of this would have been possible without my family. I am endlessly grateful to my husband William Mirone, whose care and understanding basically carried me through graduate school. I am also thankful for the support and assistance of my mother, father, and grandmother; my father-in-law, John Mirone; and the late Frances Harbour, to whom this thesis is dedicated.

Table of Contents

1 Introduction

Psychology is said to be having a 'replicability crisis'.[1] Instances of non-replication are not unusual for experimental research, but these failures are particularly disturbing against the backdrop of psychology's low rate of published replication attempts and strong publication bias in favor of positive results.[2] One can identify two closely related tasks for psychological researchers: to increase the number of replications, and to identify and combat factors that underlie poor reliability to begin with.

In line with the former task, researchers have come together to form large-scale efforts such as the Reproducibility Project: Psychology and the Many Labs Replication Project. These coordinate among many research groups in order to critically evaluate already-published effects through the process of replication. Researchers working on the latter task have proposed both sociological and methodological strategies to prevent unreliable findings from proliferating. Sociological suggestions center largely around changing the publication process (e.g., installing stricter/clearer publication standards) and even changing the overall incentive structure, while methodological ones tend to recommend bans or reforms on the use of significance testing.[3]

Though a lot of work has gone into attempting to increase the reliability of results, psychologists' strategies betray confused methodological underpinnings. Despite increased debate over replication quality and testing techniques, little headway has been made resolving fundamental conceptual issues underlying the use of statistical tests and the development, implementation, and interpretation of replicability checks. Beyond explanations of how abuses of

[1] Sometimes it is called a reproducibility crisis or replication crisis. Use is widespread, but for examples see Ritchie, Wiseman, and French 2012; Pashler & Harris 2012; or Francis 2012b.

[2] For one attempt at estimation, see Makel, Plucker, and Hegarty 2012.

[3] Suggestions for reform abound, but detailed instances can be found from the APA on Journal Article Reporting Standards (2008); Nosek, Spies, and Motyl 2012, p. 619; and Brandt et al. 2014.

the procedure called null-hypothesis significance testing – NHST – could result in findings that are difficult to replicate, these have been treated as two separate issues. This is obviously problematic: if we want the results of replication studies to warrant conclusions about research hypotheses, there needs to be a basis on which we can say that those studies are reliable.

1.1 Plan

In this essay, I argue in favor of using an error-statistical framework for addressing problems related to the replicability crisis. In the first section I introduce Deborah Mayo's error-statistical philosophy and the associated account of evidence. In sections 3-4, I show how the severity requirement can be used to guide consistent interpretations of frequentist statistical tests. This account highlights the underlying logic of tests to solve problems about their use and identify their abuses, such as the practice of inferring support for a research hypothesis directly from a rejection of the null statistical hypothesis.

I argue that psychologists' criticisms of NHST tacitly appeal to a minimal 'severity requirement': that if it is too easy for a procedure to find support for a research hypothesis H erroneously, then H fails to pass a severe test. If one accepts such a requirement, one acknowledges that the reliability of methods is important and that inquiries must have some way to account for practices that increase the probability of making an erroneous inference. The frequentist or error-statistical account is sensitive to the fact that p-hacking, verification bias, and so on corrupt error probabilities; however a Bayesian approach cannot take them into account because of its focus on comparisons as opposed to the sampling distribution. Denying their relevance becomes obviously problematic in cases where these practices enable verification bias and easy support for pet hypotheses.

In section 5, I put forward an error-statistical account of replication studies in which replicability checks discriminate genuine effects from artifacts.[4] In line with the severity requirement, results of replicability checks will only support hypotheses to the extent that that they are the result of *severe error probes.* Lastly, in section 6, I expand on the already-existing error-statistical notions of *fallacies of acceptance* and *fallacies of rejection* to show how error-statistical reasoning can be used to address two seemingly unrelated questions:

- How can negative replication results ever count against a particular research hypothesis, when it is always possible that some difference between the replication and the original experiment could have blocked the effect from obtaining?
- What are we licensed to infer when one or more direct replications finds statistically significant results, consistent with the findings of the original experiment?

[4] By genuine effects, I mean something like reliably inducible phenomena; although I do not give an account of this here, effects might include everything from estimates of parameters, to correlations between a set of variables, to specific function forms. This very broad definition hopefully captures the variety found in scientific hypotheses.

2 Background

Deborah Mayo's error-statistical philosophy of induction is rooted in the idea that for a hypothesis to be warranted, it must survive a severe test. Under this account, the evidence for a hypothesis is not described by a simple function relating the hypothesis and the data, but is also dependent on features of the test that the hypothesis was subjected to.[5]

To clarify, it seems obvious that for data to provide good evidence for a hypothesis, it is necessary for it to accord with that hypothesis in some way. However, *mere* accordance between the data and the hypothesis – on whatever notion of accordance one is interested in – is not *sufficient* for there to be good evidence. Such a loose criterion would leave us highly vulnerable to certain forms of bias; for example, one can imagine a case where a scientist's experimental apparatus malfunctions in such a way that it only generates data consistent with her pet hypothesis. The observations will be consistent with her hypothesis, but it would be strange to say they provide good evidence for it.

In the error-statistical account, Mayo captures this intuition in the form of a **severity requirement**: that for a hypothesis to be supported by the data, it must pass a **severe test** with that data (1996, 180). Whether the test is severe is determined by the following rule:

> **Severity criterion:** Test *T* severely passes hypothesis *H* with data *e* iff: [6]
>
> a) *e* accords with (fits) *H,* and
> b) *T* would, with high probability, have yielded a result that fits less well with *H* than *e,* in a condition where *H* was false. (Equivalently: It is very improbable that *T* would have yielded a result that fits about the data generation or as well with *H* as *e,* under the condition that *H* were false.)

[5] For examples, see Mayo 1991, 526–528; Mayo and Spanos 2006; Mayo and Cox 2006.
[6] From Mayo 1991, 528.

In some cases (e.g., significance tests), the relevant probabilities can be strictly quantified. At other times, our understanding of severity must be more loosely understood. If strict error control cannot be guaranteed, one must still appeal to known sources of error and bias to rule them out as strongly as possible, or make use of inferences that are robust even in the presence of unknowns. One can think of this as a minimal normative recommendation.

This account can be usefully contrasted with Karl Popper's account of severe testing. For Popper, there was no method that allowed scientists to infer reliable inferences from observations to hypotheses; as a result, such inductions could not ever be justified (1953, 53). In that case, science – instead of being a process of drawing inferences – was a process of falsifying hypotheses. Hypotheses are corroborated when they survive severe testing, which must subject them to high risk of refutation; however, this corroboration means nothing more than the history of failed attempts to refute the hypothesis (36). For Popper, the only thing we learn when a hypothesis survives testing is that there is still a possibility that it could be true.

The error-statistical account, while still emphasizing severe testing, denies this with the contention that we *can* learn something about the hypothesis when it survives testing. If a particular test has a high probability of detecting an error when it is present – such as the bias caused by a confounding factor that could also account for these results – and it passes the hypothesis nonetheless, there is support for inferring that the error is, in fact, absent. This is because, with high probability, the procedure would have flagged that error or rejected the hypothesis were it present.

2.1 Calculating severity for significance tests

Significance tests, like all frequentist analyses, are performed in the context of a statistical model of a data-generating process. This model gives the set of all events in the sample space, along with the distribution of probabilities that have been assigned to those events

according to statistical hypotheses in the model. These hypotheses provide mapping rules that allow inferences from the characteristics of the sample to model parameters (Mayo & Spanos 2011, 154-155).

The first component of a significance test is the specification of a null hypothesis about some unknown parameter – for example, the difference between the mean response measurements for two groups. A researcher interested in whether a treatment group scores higher than a control group on a particular measure might specify the null hypothesis H_0 as:

$$H_0 : \mu_1 - \mu_2 = 0$$

That is, that there is no mean difference between groups μ_1 and μ_2. This is an example of what is colloquially known as the *nil hypothesis:* a null hypothesis that claims there is zero difference between groups, or zero correlation between variables.[7] This hypothesis is accompanied by the (sometimes only implied[8]) directional alternative hypothesis:

$$H_1 : \mu_1 - \mu_2 > 0$$

The next component is the significance level (α-*level*) of the test. This is the probability α of making a type-1 error: that is, rejecting the null hypothesis on the condition that the null hypothesis is true. The α-level specifies a cutoff c_α beyond which H_0 will be rejected.[9]

Thirdly, one must select a test statistic *d(x)* whose distribution is known under H_0 (and ideally under alternative hypotheses). The observed difference in means (or the value of any sample statistic) will be translated into the test statistic by putting the value in terms of the

[7] This term was originally coined by Jacob Cohen – see (Cohen 1994, 1000).

[8] While a Fisherian significance test only specifies a null hypothesis to be rejected or not rejected, significance tests in psychology seem to have implied alternative hypotheses. Without a directional alternative hypothesis, the notion of statistical *power* is nonsensical; the alternative is similarly needed to calculate severity.

[9] This is common knowledge; for discussions of the related reasoning, see (Mayo & Cox 2006) and (Gelman and Loken 2013).

standard deviation of the probability distribution expected under H_0.[10] The threshold c_α, which delineates the rejection region for the test, is also put in these terms.

Consider the example of an independent-means t-test. By randomly assigning subjects to treatments, we are able to compute the probability of finding a difference as large as the one observed between $\bar{x}_1$ and $\bar{x}_2$ merely by chance. (If there is no statistically significant effect, then the observed difference will be likely due to arbitrary assignment.) When the null hypothesis claims the difference between the means of two groups is less-than or equal to zero – in other words, that the treatment has no positive effect – the conversion of the sample statistic into the test statistic is given by the following equation, where $\bar{x}_1$ and $\bar{x}_2$ are the means of the treatment and control group, n_1 and n_2 are their respective sample sizes, and s_p is their pooled standard deviation:[11]

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

If each sample has 50 participants, there will be 98 degrees of freedom; using this value, one would look up the t-distribution and see that the critical value at $\alpha = .05$ is $t \approx 1.66$.[12] This is the threshold for rejection. Our test rule, for at $\alpha = .05$, is therefore:

Whenever $\{T > 1.66\}$, reject H_0.

The test procedure determines how much the sampling distribution departs from the distribution expected under H_0, and the p-value gives the probability that a difference greater than

---

[10] The general formula for constructing such a distance measure $d(\boldsymbol{x})$ is the following:

$$d(x) = \frac{\bar{X} - \mu_0}{standard\ deviation}$$

[11] The example used in the text draws on (Leary 2004, 262–266) for instructions on constructing a two-sample independent t-test, specifically a *pooled* t-test (which assumes variances are equal).

[12](Filliben 2003)

or equal to the outcome would be obtained under the null hypothesis using this sampling method.[13] But what does it mean for H_0 to be rejected – for a particular outcome to be significant at $p < .05$? Do we have warrant to infer a particular hypothesis?

[13] See (D. G. Mayo and Spanos 2011, 155–156)

3 Threats to the interpretation of experimental results

The correct interpretation of statistical tests, and the potential role of statistical significance testing in the propagation of unreliable findings, is an issue that comes up repeatedly in the replicability discussion. The binary outputs of significance tests – rejecting H_0 or failing to reject H_0 – are conducive to making erroneous inferences. The actual properties of the tests also make them sensitive to certain sources of error. Accordingly, misunderstandings about these properties – as well as abuses of them –are widely believed to have contributed to the replication crisis.[14]

Several risks can be identified in the interpretation of significance tests and associated analyses:

- The actual significance level of a particular outcome may not be reflected by the computed significance level (due to p-hacking, etc.)
- Even when the above is not the case, a statistically significant result does not necessarily indicate an effect of a magnitude large enough to be meaningful; and insidiously,
- A (legitimate) statistically significant result does not necessarily support the corresponding research hypothesis.

These problems share a common element: they involve an inference to a hypothesis that has not been severely tested. In the following section, I will explain the first two hazards described and summarize some of the criticisms associated with them. After that I will focus will be on the third point, which seems to be the one that has been most neglected by the discussion. The upshot is that if you accept the severity requirement, certain principles for interpreting the output of statistical tests will follow, that allow us to avoid making particular errors.

[14] For just a few articles attributing non-replicable results to the use of significance tests, see Lambdin 2012; LeBel and Peters 2011; Wagenmakers et al. 2011; Galak et al. 2012; and Gelman and Loken 2014.

3.1 The computed p-value may differ from the actual p-value

The p-value calculated for a particular outcome will only accurately represent the error probability when certain experimental assumptions are met. Practitioners frequently violate these assumptions in ways that inflate the statistical significance of results (Simmons, Nelson, and Simonsohn 2011). These methods include using multiple comparisons without reporting them, data-peeking, performing (but failing to report) any non-significant statistical investigations, and a number of related actions - all of which increase the risk of finding some significant result or other even if the tested hypotheses are all false (Simonsohn, Nelson and Simmons 2013, 534).

These significance-altering practices were prominent in the criticisms of Daryl Bem's famous *psi* research, in which he purported to show evidence of precognition in 8 out of 9 studies. According to his critics, Bem's data was biased from presenting exploratory research as confirmatory research (Wagenmakers et al. 2011). In addition to just testing his pre-specified hypotheses, he tested additional hypotheses based on any promising-looking interactions in his data-set. The significance threshold should have been adjusted to compensate for this, but it was not (ibid., 427).

The consequences are problematic regardless of whether researchers like Bem intend to massage their data into providing fraudulent results. As Andrew Gelman and Eric Loken point out, these significance-level-enhancing techniques are often unintentional, and can affect the error probabilities in ways that would be difficult to except without understanding the underlying logic of the test. For example, when the specifics of data selection and statistical analysis are not specified in advance, the appearance of the data collected often dictates which analyses are ultimately run (Gelman & Loken 2014, 1-6). By increasing the probability of rejecting the null hypothesis through multiple chances, the test procedure has an increased probability of providing data consistent with the alternative hypothesis, *regardless* of its truth status. The correct

computation of the p-value will now represent one's probability of observing *any* significant differences.

To respect the severity requirement, the researcher must audit her results to see if the assumptions of the statistical test have been correctly met. If certain assumptions have been violated, the computed p-value will be different from the actual p-value. Even in cases where the computed value is accurate, however, additional problems can arise. For example, a statistically significant result does not necessarily indicate a meaningful effect – hence results require post-data severity analysis.

3.2 A departure from the null hypothesis may be mistakenly interpreted as 'more significant' than it actually is

A statistically significant result does not necessarily indicate a meaningful effect size. That is, it may be the case that a discrepancy from the null hypothesis is genuinely statistically significant, but that the magnitude of the effect is not as great as the one suggested by the rejection. Following Mayo and Aris Spanos, this sort of mistaken inference will be termed a *fallacy of rejection* (2006, 341–345).

This rejection fallacy is related to one of the biggest criticisms lobbed at significance test usage in the replicability discussion. The charge is that p-values fail to provide insight into the evidential value of results, because whether or not the null hypothesis is rejected seems to merely reflect the size of the sample. If you hold the size of the test α constant, then as n increases, so does statistical *power:* the capacity of the test to detect a discrepancy from the distribution expected under H_0 in the direction of the alternative hypothesis, or P(Reject H_0; H_1 is true). With the way the test is designed, this means there are sample sizes large enough for the test to generate a statistically significant result in cases where there is just a small or trivial underlying discrepancy from H_0 (Mayo & Spanos 2011, 174).

If, as the philosopher and psychologist Paul Meehl has argued, the null hypothesis is always technically false in the social sciences,[15] then whether or not H_0 is rejected will merely be a matter of how large *n* is (1978, 822). Meehl observes that in psychology, increased test sensitivity (i.e., power) is associated with a decreased risk of falsification. Rejection of the null hypothesis is routinely taken to license an inference to the research hypothesis; but because it is so easy to get a significant result with a large sample, rejecting H_0 under that condition can only offer "feeble corroboration" for the research hypothesis (*ibid.*).

Recall the severity criterion laid out in Chapter 2. When H_0 is rejected in the sort of large-*n* cases described above, part (a) of the criterion is met because the rejection of H_0 involves accordance with an alternative hypothesis consistent with the research hypothesis. Part (b), however, remains to be fulfilled: the increased test sensitivity must be taken into account by the researcher, to prevent researchers inferring from an effect of a larger magnitude than is actually warranted.

The severity interpretation of results dissolves this problem by directing us to determine just how large of a discrepancy from H_0 we are warranted to infer from the observed outcome (Mayo & Spanos 2011, 174).

Suppose we are testing the following hypotheses:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

Recall that running a significance test amounts to checking how much the sampling distribution departs from the one that would be expected under H_0, and the p-value is the

[15] This should be understood as a metaphysical claim about the relationship between the objects of investigation in psychological research; see Meehl 1978, 822). I will not evaluate it here, and doing so is not necessary for responding to the concern about the use of the nil hypothesis.

probability that a difference greater than or equal to $d(x_{obs})$ would be obtained under H_0. When we set a low value for α, we in effect are establishing in advance that whenever the test rejects H_0 at $d(x_{\text{obs}}) > c_\alpha$, H_1 passes with high severity. Here, the value demanded by the statistical version of the severity requirement is the probability of getting $d(x_{\text{obs}})$ or a more extreme value if the null is true; the severity with which H_1 passes is therefore calculated as

$$SEV((\mu_1 - \mu_2) > 0) = P((d(x_{\text{obs}}) \geq d(X));\ (\mu_1 - \mu_2) \leq 0)$$
$$= 1 - p,$$

where p is the statistical significance level of x_{obs} [see Appendix A for an example].

We are still at risk of making a fallacious inference, however. It is true that *some* discrepancy is warranted, but this is very general information; we have not learned whether or not the discrepancy is merely trivial. To warrant the inference to an effect of a particular magnitude – some discrepancy δ such that H': $\mu_1 - \mu_2 > \delta$ – we must calculate the probability of getting an observation as or more extreme than $d(x_{\text{obs}})$ in the event that H' were not the case:[16]

$$SEV((\mu_1 - \mu_2) > \delta) = P((d(x_{\text{obs}}) \geq d(X));\ (\mu_1 - \mu_2) \leq \delta)$$
$$= P((d(x_{\text{obs}}) \geq d(X));\ (\mu_1 - \mu_2) = \delta)$$

If there is a high probability of getting as large an outcome as $d(x_{\text{obs}})$ under the condition that $(\mu_1 - \mu_2) = \delta$, then the severity with which $(\mu_1 - \mu_2) > \delta$ passes is low.

The severity level limits the effect size that one can responsibly infer with a high-powered test, as for any particular effect size severity will be inversely correlated with the size of the sample.

3.2.1 Confidence intervals and severity

[16] Here I follow demonstrations from Mayo and Spanos (2011), 168–174.

The severity requirement directs us to report the differences in means that the observed effect would be improbably far from, given the null hypothesis. These values are well-captured by confidence intervals. Say we are interested in estimating the difference for a variable between two groups μ_1 and μ_2. The variance of the parameter of interest is unknown, but assumed the same for both groups. Suppose further that we want to make a $(1-\alpha)\%$ confidence interval, where $\alpha = .05$. The independent samples $n_1$and n_2 each have 50 subjects, meaning there are 98 degrees of freedom to take into account.

At $\alpha = .05$ we would calculate[17] the interval as follows:

$$.95 = \mathrm{P}\left(\left([x_1 - x_2] - 1.984\left(\mathrm{SE}_{x_1-x_2}\right)\right) \leq (\mu_1 - \mu_2) \leq \left([x_1 - x_2] + 1.984\left(\mathrm{SE}_{x_1-x_2}\right)\right)\right)$$

We collect our data and find that $\bar{x}_1 = 30$ and $\bar{x}_2 = 27$, and $\mathrm{SE}_{x_1-x_2} = 1.2$. Then our confidence interval will be:

$$.95 = P\left(\left(3 - 1.984(1.2)\right) \leq (\mu_1 - \mu_2) \leq \left(3 + 1.984(1.2)\right)\right)$$

$$.95 = P(.612 \leq (\mu_1 - \mu_2) \leq 5.381)$$

Given our assumptions, 95% of the estimated ranges generated with this sampling method would contain the true value of the difference between means. Now consider the following hypotheses:

$$\mathrm{H}_1 : (\mu_1 - \mu_2) > .612$$

$$\mathrm{H}_2 : (\mu_1 - \mu_2) < 5.381$$

Both H_1 and H_2 pass with severity equal to the confidence level (.95), meaning we have good warrant for inferring each one. If H_1: $(\mu_1 - \mu_2) > .612$ were not the case, then it would have been highly probable for us to have found a smaller difference in means than we obtained;

[17] T-values from *NIST/SEMATECH e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, 5/4/2015.

and if $H_2 : (\mu_1 - \mu_2) < 5.381$ were false, then with high probability we would get a *more* extreme difference in means.

3.2.2 Unwarranted inferences

In addition to reporting warranted inferences, taking the error-statistical account seriously demands we examine which effects are *not* well indicated by the data – this information "points to the direction of what may be tried next," as Mayo and Spanos write, "and of how to improve inquiries" (2006, p. 346). Appropriately, then, one offering a severity analysis of their data must also provide an upper bound show those discrepancies from H_0 that would be unwarranted. This directly addresses the fallacy of rejection, by determining if only trivial differences have passed the test severely. A good benchmark to test is whether the population mean exceeds the sample mean, or in our case:

$$H_B: (\mu_1 - \mu_2) > (\bar{x}_1 - \bar{x}_2)$$

One can report such results alongside confidence intervals, as important supplementary information. However, even when it has been determined that we have evidence for a significant effect of a specific magnitude, one must be wary of committing an additional fallacy of rejection. This is because determining whether an inference from the statistical hypothesis to the *substantive* hypothesis is warranted requires an examination of parts of the inquiry beyond the statistical model.

4 Rejecting H_0 does not guarantee support for the research hypothesis

A detailed account of how to interpret the results of statistical tests is essential to confidently addressing the replication crisis, but psychologists are remiss to focus exclusively on statistical aspects of the problem. One of the biggest potential hazards concerns researchers' tendency to interpret rejection of the statistical null hypothesis as acceptance of one's preferred *substantive* hypothesis. The substantive hypothesis, in the words of Paul Meehl, "is the theory about the causal structure of the world, the entities and processes underlying the phenomena" (1978, 824). Meehl criticized the leap from rejecting the null hypothesis to inferring the substantive hypothesis, arguing that from a Popperian perspective such an inference goes far beyond what we can reliably conclude from rejecting the null (*ibid.*).

Similar criticisms have been put forward by other researchers, though the reasoning behind them varies. Etienne LeBel and Kurt Peters (2011) argue that this inference is problematic because of researchers' use of the nil hypothesis, which they call a "straw man—a bit of statistical fluff with no theoretical substance" (374). The nil hypothesis has been claimed to be always false in the social sciences, and in addition its specification is unmotivated by theoretical content. Therefore, the criticism goes, its rejection does not seem to tell us about what alternative hypothesis we should infer (*ibid.*).[18]

However, the severity construal of tests – which takes statistically significant results as evidence for a particular effect size – immediately circumvents this problem. It is not problematic for the nil hypothesis to be an artificial construction if its purpose is to show what discrepancies from that distribution we would be warranted to infer (recall section 3.2).

[18] Charles Lambdin goes so far as to say that the artificiality of the null makes getting a p-value below the significance threshold nothing but an "easily achievable expression of baloney" (2012, 80-82).

Others criticize the leap based on the fact that significance tests do not take into account prior probabilities. In their commentary on Daryl Bem's psi studies, Wagenmakers et al. argued Bem was mistaken to infer support for psi from statistically significant results, because this ignored how improbable it would be for precognition to exist. After assigning a low prior probability to the existence of psi (their description of H_1), they proceeded to show that even if a confirmatory result were many times more likely under H_1 than H_0, the posterior probability wouldn't expect to approach a point where H_1 became probable or choiceworthy.[19] Such a result could come about with a larger sample size:

> "Based on the mean and sample standard deviations reported in Bem's (2011) Experiment 1, it is straightforward to calculate that around 2,000 participants are sufficient to generate an extremely high Bayes factor BF_{01} of about 10^{24}; when this extreme evidence is combined with the skeptical prior, the end result is firm belief that psi is indeed possible."[20]

Note Wagenmakers et al. consistently discuss the appropriate competing statistical hypotheses as hypotheses of precognition existing versus not existing. This strategy seems to make the same mistake they are criticizing Bem for making, only dressed in Bayesian clothing. There's a reason to understand statistical hypotheses H_0 and H_1 as describing ranges of values for outcomes or parameters, that should motivate frequentist and Bayesian critics alike: whether psi exists is still a different question from whether or not one's experiment generates data consistent with psi existing. For example, one can imagine a graduate student sabotaging experiments so that the computer subliminally informs participants the right answers for their activity, causing an increased frequency of correct predictions. If this were the case, we would not want to use an inference tool that would dismiss the existence of *any* effect due to the low prior probability for *psi* – the effect is real, and important for us to know about. It's just not the result of precognition.

---

[19] (Wagenmakers et al. 2011, 429)

[20] (Wagenmakers et al. 2011, 430)

A related point should not be overlooked. The skepticism of Wagenmakers et al. reflects the low rate of paranormal belief among psychological researchers (and other scientists); under the recommended analysis, these researchers will want to assign a low prior probability to psi hypotheses, presumably because there is no known mechanism by which it could plausibly work. Still, as Bem points out, "the discovery and scientific exploration of most phenomena have preceded explanatory theories, often by decades or even centuries" (2011, 408). We have upheld false scientific beliefs in the past, so by what right should we allow the *apparent* plausibility of motivating theories to dictate the evidential import of observations? Wagenmakers et al. show that conflating levels of inference, or jumping from the statistical to the substantive, is not a hazard that is unique to significance testing.

If the problem is not reducible to any of the criticisms above, why is it that the substantive leap is so dangerous? Up until this point, we have considered severity reasoning in a very narrow sense. To explain why the substantive leap is so problematic in psychology, we must explore a more general notion of severity.

4.1 Probing substantive hypotheses with general severity reasoning (GSR)

The argument from a severe statistical test is a type of *argument from error*. Arguments from error draw on the following reasoning:

> **Argument from error:** If our data accords with hypothesis *H* and results from a procedure that, with high probability, would have produced results more discordant with *H* were *H* incorrect, then the correctness of *H* is indicated (Mayo 1996, 445).

This type of argument warrants inferences in line with a general severity requirement:

> **GSR:** The hypothesis that an error is absent is warranted by the data if, and only to the extent that, the inquiry procedure 1) had a high probability of revealing an error were it present and 2) did not detect such an error (Mayo 1994, 273).

On levels beyond statistical reasoning, where inference often requires juggling informal probabilities, the severity requirement becomes a demand for tests with the capacity to detect

error. Data only warrant a claim to the extent that it would not be easy to find such a good fit between *H* and *e* even in a world where *H* was false.[21]

Procedures with a high capacity for finding errors are called "severe error probes" or, in more formal contexts, “severe tests” (1996, 7). In simple cases, they may consist of a single test, but in others they can be built out of several smaller inquiries; severely probing a hypothesis typically requires breaking inquiries down into manageable pieces, such as checking experimental assumptions. The error-statistical account will treat replicability checks as these kinds of inquiries, and judge them as legitimate or illegitimate to the extent that they subject hypotheses to severe tests.

The error-statistical account (1996) asserts that although scientists must sometimes try to arbitrate between high-level theories, most of scientific work, especially experimentation, is concerned with more local goals – particularly the discrimination of genuine effects from artifacts and the identification of backgrounds.[22] Inquiries will be interested in determining the values of parameters, the degree to which effects vary, how particular confounding factors can interfere with instrument readings, and a whole host of other ground-level questions. Statistical testing tools have an extremely important role to play in this picture, as they can be used to determine the probability of getting particular results if a particular hypothesis is true; this allows us to gauge how well data accord with the hypothesis, as well as how probable that data would be if its complement were true.

---

[21] I should emphasize that claiming that a hypothesis is warranted is different from claiming the hypothesis is true. When we claim that hypothesis is warranted, we are saying we have grounds for accepting it as reliable, or thinking that the effect it describes approximates what can be brought about consistently in the experimental context. The amount to which one is actually warranted in accepting such a hypothesis is proportional to the strength of the argument from error; in informal cases, the burden of proof is on the researcher to show that they have adequately controlled the experiment.

[22] (Hacking 1983, 171–176; D. G. Mayo 1994, 270–272)

From an error-statistical perspective, the problem with inferring support for the substantive hypothesis from statistically significant results is that the significance test (ANOVA, regression, etc.) does not probe for errors on the substantive level. The rejection of the null may, ultimately, support such an inference, but only when we have shown that the research hypothesis is connected to the model of the data in the right way. This requires us to rule out sources of error such as problems with experimental design, experimental assumptions, and model specification.

4.2 Connecting the research hypothesis and the statistical hypothesis

The error-statistical account views inquiries as involving multiple levels of models and hypotheses. Mayo associates these levels with different corresponding error threats: in addition to trying to determine if the research hypothesis is false, researchers must rule out such possibilities as that the extent of the hypothesis was erroneously estimated, that the factors were modeled incorrectly, that the assumptions of the statistical model were false, and so forth (1996, 140).

For severe testing, hypotheses should be constructed so that their denials are informative in a specific way (190–191). To illustrate, take the following "primary level" research hypothesis:

> **H_0:** The difference between the treatment group and control group is predicted by chance: $\mu_T - \mu_C = 0$.

Notice that the falsity of such a hypothesis would imply the alternative hypothesis,

> **H':** The difference between the treatment and control group is greater than the amount predicted by chance: $\mu_T - \mu_C > 0$.

The way this is structured, it may seem immediately amenable to a statistical test. However, this research hypothesis cannot be directly related to the recorded data without an additional model standing in between. We must connect the hypothesis to the observations that

are made. Researchers accomplish this (not necessarily explicitly) through the use of an experimental model, which creates a route by which data can weigh in on higher-level questions of interest.

Severe tests must be understood within a piecemeal context, where parts of the inquiry are linked to one another and errors are ruled out at several levels. Importantly, if the data model is not adequately linked to the primary model, *then we will not be warranted in making an inference from the results of the statistical test.* If the experiment itself is designed poorly – a simple example would be the use of a bad proxy variable to measure the parameter of interest – then even if our significance test rejects the null hypothesis, we won't be warranted in inferring the research hypothesis is correct. If the experiment fails to genuinely measure what it claims to, it might find significance for a wide variety of reasons, even when the research hypothesis is utterly false. (For example, unscrambling age-related words may not have any priming influence on walking speed, but observational bias on behalf of unblended research assistants may yield the appearance of such an effect nonetheless.)

5 Error-statistics and replicability issues

As suggested by the previous sections, criticisms of significance test use focus on how easily these tests be used to find support for a preferred hypothesis. This seems at odds with the replicability crisis: if it is so easy to find support with these tests, then why would researchers have such a difficult time replicating the results of earlier studies? The critic needs to be able to reconcile these hypotheses.

When researchers attempted to replicate the results of Bem's studies on retroactive interference, they were in effect using pre-designated hypotheses – the analyses were partially constrained by the ones reported in the original study. Bem, in contrast, had considerable freedom to decide which analyses he was going to run, and maybe even which ones he was going to report.[23] This reflects a more general pattern in which authors of the original studies may have analyzed their data using techniques that increase the probability of rejecting H_0, but the replicators are more constrained. They typically base the decision of which hypotheses to test and comparisons to run off of the analysis that was originally reported, and in the case of replicability projects, the analyses are also pre-registered.[24]

If the problem with unreliable findings is that researchers are running multiple comparisons without correction, cherry picking, and so on, then there is an obvious motivation for preferring a framework that allows us to account for the ways error probabilities are altered by these kinds of practices. Interestingly, the alternative statistical techniques being promoted by critics of significance tests do not have a way to take this into account. Unlike significance tests, posterior probabilities and Bayes factors are not affected by the sampling distribution of a

[23] For an example and discussion see Galak et al. 2012, especially p. 943.

[24] Registered confirmatory analyses plans are reported for both the *Many Labs Replication Project* and the *Reproducibility Project*; see Klein et al. 2014, 147 and Aarts et al. 2013, 13 respectively.

particular test statistic. Bayesian approaches rely on the assumption that the only information in a sample that is relevant to the parameters of a model is the information in the likelihood function.[25] The sampling distribution is not relevant to inferences – meaning error probabilities, which depend on the sampling distribution, are not taken into account. As a result, error-probability –altering considerations such as stopping rules are "irrelevant" (Berger and Wolpert 1988, 74).

These practices do not affect Bayes factors or posterior probabilities, but they absolutely can trick the associated tests into showing support for erroneous inferences. Simonsohn (2014) recently demonstrated, for example, that selective reporting and other p-hacking techniques increase the probability of making a Type-I error using these tests (5-9).[26]

It is not clear why switching to these analyses, then, would alleviate the concerns associated with significance test use.

5.1 An error-statistical account of replication

Criticizing NHST on the grounds that its misuses make it too easy to get confirmatory results implies tacit agreement with the following reasoning: a procedure is not a good source of evidence if it will consider almost any observation to be support for *H.* If a researcher agrees with this, she should also agree to certain interpretations of replication results.

In the following sections, I will examine questions about positive results in the case of replication studies, and examine a problem that has gone grossly overlooked. In order for this discussion to make sense, however, I will need to introduce the concept of replicability and describe how it fits into the error-statistical account.

---

[25] See (D. G. Mayo 1996, 339–341)

[26] Also see Mayo 1996, chapter 10.

5.2 Replication, at first glance

For the sake of this discussion, replicability will be defined as the ability of a particular effect to be brought about consistently.[27] High replicability is *prima facie* support in favor of a hypothesis while low or no replicability is either no evidence or evidence against it. The process of replication is performed to discriminate experimental effects from artifacts; according to researchers from the Reproducibility Project,

> "Independently reproducing the results [of an older experiment] reduces the probability that the original finding occurred by chance alone and, therefore, increases confidence in the inference. In contrast, false positive findings are unlikely to be replicated." (Aarts et al., 2013, 3)

Freak coincidences, questionable statistical practices, and the like can create the illusion of support for a hypothesis, but at first glance it is unlikely that an experimental finding that has been brought about over and over again will be the result of such errors. Artificial findings should be difficult to bring about consistently, in comparison to outcomes reflecting a genuine effect; at the very least, some sort of systematic cause would be implied.[28]

Replicability suggests something about an experimental or statistical outcome: that it is in some sense genuine or reliable.[29] Here the error-statistical account follows Popper and Fisher[30] in the position that one cannot take a result to be genuine unless it could be replicated, since flukes are not going to be reproduced on command. (The same does not hold, of course, for findings that are the result of systematic biases.)

---

[27] This is loosely adapted from the Reproducibility Project: Psychology, who define reproducibility as "the extent to which consistent results are observed when scientific studies are repeated" (Alexander et al., 2012, 3). The terms reproducibility and replicability are used interchangeably in psychology, but for convenience I will stick to replicability.

[28] This view can also account for replications meant as fraud checks – there too, however, one is attempting to rule out a biasing factor.

[29] By genuine effects, I mean something like reliably inducible phenomena; although I do not give an account of this here, effects might include everything from estimates of parameters, to correlations between a set of variables, to specific function forms. This very broad definition hopefully captures the variety found in scientific hypotheses.

[30] See Popper 2005/1959, 23; Fisher 1926, 504.

Replications are used to mount arguments from error. Although they can be broken up into different categories, they really exist on a continuum of closeness to the original study. **Direct replication** is the process of repeating a scientific experiment using the same procedure and data analysis as the original study.[31] This process is typically contrasted with **conceptual replication**, which uses a different procedure from the original study to see if the same effect can be brought about using a different method (Aarts et al., 2013, 2). If we accept the severity requirement, then we should assess these inquiry methods based on how well they severely probe for error and allow us to learn from testing. The force of the argument from error to a particular inference will depend on (1) the probativeness of the studies[32] that compose the 'sub-inquiries' and (2) the probability of getting the results that they did if the hypothesis were false. *Which* hypothesis passes depends on the sort of variability that exists between the original study and the replication(s), and the sort of errors they are able to rule out.

5.3 Cartwright's account

Philosophy of science shows a surprising paucity of work on replicability specifically. When they are not referring to Popper, psychologists often reference Nancy Cartwright, one of the new experimentalists mentioned earlier, for her account of replication sketched in a response to Harry Collins.[33] There she draws a distinction between two different kinds of replicability beyond basic checks of proper implementation, *replication*, and *reproduction*, the latter two being somewhat analogous to the way psychologists use the phrases *direct* and *conceptual replication* (Cartwright 1991). Under her system, replication is used to check that a particular

[31] Aarts et al. 2013, 2; Earp and Trafimow 2015, 11–12

[32] Studies do not have to report results in this format, but I will be exploring simple cases of a kind typically observed in psychological research projects.

[33] Discussing replication methodology, psychologists refer to Collin's work quite frequently – which is confusing, considering he only offers a sociological account of replication. See Collins 1984, along with Alexander et al. 2012 and LeBel & Peters 2011 for examples.

instrument is operating 'correctly'. When it is, applying that instrument[34] to test a hypothesis is supposed to result in a particular outcome. An extremely close replicability test will be limited to checking that a particular experimental procedure yields these results consistently (*ibid.* 146).

Reproductions, on the other hand, are used to determine if an effect exists beyond the narrowly defined conditions of a particular experiment, by using different instruments and methods to test the same hypothesis (149-151). Reproducibility protects inferences from problems with experimental instruments via an argument from coincidence (150). The following is a reconstruction of such an argument:

1) Suppose that a set of *n* experiments with varying designs $\{p_1, p_2, \ldots p_n\}$ each test $\mathcal{H}$ and report finding e.
2) Further, assume that $\{p_1, p_2, \ldots p_n\}$ from (1) are well-designed and executed, and that they draw on different and *independent* assumptions from one another.
3) It is possible for any individual p_x to report an erroneous outcome.
4) Under the conditions in (2), it would be a notable coincidence if $\{p_1, p_2, \ldots p_n\}$ all reported mistaken results that all happened to be e.
5) The truth of $\mathcal{H}$ is the most likely cause of the agreement about e (154).
6) [Implicit premise] Of a set of candidate hypotheses for some piece of evidence, one should accept the one that is the most likely cause.[35]

C) We should accept $\mathcal{H}$.

This argument has a similar structure to the argument described in section 0, though it appears to depend on abduction.[36] For now, let's put aside the problems courted by inference-to-the-best-explanation –type argument and note that even if we grant Cartwright's reasoning, it's not clear that it would support a strong inference in the context of psychological science. Premises (4) and (5) are especially suspect, simply because the massive variety of invisible factors assumed to be at play. In a set of conceptual replications, even quite different

[34] 'Instrument' should be understood broadly here, to include anything from a particular device to an entire experiment.

[35] Cartwright here is relying on an *inference to the most likely* or *best cause (IBC),* which takes to be importantly different from *inference to the best explanation.* See Psillos (2008) for a critical discussion of the relationship between IMLC and IBE.

[36] In an IBC argument, one infers $\mathcal{H}$ because $\mathcal{H}$ being true is the most likely cause of *e* (Cartwright 1991, 153-154).

experiments will be forced to rely on shared experimental assumptions; the threat of systemic error looms especially hard here, especially if psychology shares the sort of complications Cartwright suggests economics must contend with (151).

An error-statistical account can deal with this. In *Error and the Growth of Experimental Knowledge,* Mayo argues that a group of arguments from coincidence similar to the one Cartwright describes are instances of a more general argument pattern – the argument from error described earlier in this essay (1996, 65-67). Cartwright's argument can be re-constructed and supplemented in accordance with error-statistical meta-methodology, with the argument from error doing the same work as Cartwright's abduction. I will do this in the next section, but first I will describe how replications are used in psychology.

5.4 Direct and close replications

In a direct replication, an experiment is repeated exactly the same way as in the original study, but with a different sample. This notion is, of course, ideal. Those attempting direct replications would more realistically recreate a study procedure step-by-step according to the *Methods* section of the original published study, preferably with assistance from the authors to resolve any unclear instructions.

Cartwright asserts that the purpose of such a procedure is to determine whether a particular experiment gives consistent results when used to test a particular hypothesis (1991, 146). This is like checking if an instrument works correctly. In psychological contexts, however, that reliability question is typically addressed when designing psychological scales or performing pilot studies in general. The replication researchers certainly do not present their work as aiming at such a goal. Their activity would be better described as attempting to determine whether a particular outcome was the result of chance variation or some other source of error.

However, strict direct replications are sometimes taken by psychologists to indicate that an effect is real *independent of the instrument used to detect the effect or particular set of experimental conditions*.[37] This is beyond the scope of what a direct replication *per se* can provide strong evidence for. Depending on the design of the original experiment, for example, one might have a set of results where apparent agreement with the hypothesis is entirely attributable to an experimenter expectancy effect.[38] This refers to a phenomenon in which researchers unintentionally bias their observations to be consistent with the hypothesis they think is true. Although there are easy ways to prevent experimenter expectancy effects, in the case where the original experiment neglected to control the errors an *exact* direct replication cannot by itself shed light on whether this contributed to the effect. *Without a well-designed experiment*, the only hypothesis that demonstrable replicability will pass severely is:

$\mathcal{H}_{exp}$: Experimental procedure i reliably yields effect e.

The above inference is clearly limited, but the argument to it can be informative for a variety of tasks. One example would be for fraudbusting. When attempting to validate earlier results, one is interested in the question of whether i has a high probability of being able to result in findings like those found by the previous study. High fidelity will sometimes be essential to severely test this hypothesis; comparison of the original results with replication findings will tell us little about the validity of the original results if you are testing procedures you would not reasonably expect to observe the same outcome.

If more variation is allowed, the replication studies can become more informative with regard to normal research hypotheses. LeBel and Peters claim that close replications are

[37] (Francis 2012a, 585; Francis 2012b, 975; LeBel and Peters 2011, 375–376; Aarts et al. 2013, 4)
[38] (Leary 2004, 217–218)

preferable to conceptual replications because any changes to the methodology are minimal and *differences that do occur are intentionally introduced* (2011, 376). These changes correspond to different kinds of experimental constraints and potential errors that must be ruled out, if we want to interpret a negative replication result. The benefit of this increased control is the increased ease of assigning blame when a replication study fails.

When we are testing to see if there is warrant for inferring the 'substantive' or research hypothesis – as opposed to a hypothesis about the performance of a specific experimental algorithm – the error-statistical meta-methodology directs us to find ways to rule out the possibility that e is attributable to the particulars of a particular experimental algorithm. Here less-faithful replications can play an essential role in ruling out errors at the experimental level. The differences between close replications and the original can be planned to see if a significant effect is still obtained, thereby supporting inferences to an effect that is not dependent on such particulars. Consider the following argument: For a single close replication, the ability to replicate experimental findings in this manner might be used to argue for the local hypothesis that:

$\mathcal{H}_{\text{f}}$: Effect e is not an artifact of auxiliary factor x.

If e is not observed when $\mathbf{x}$ is removed, or if variation in e can be attributed to variation in x, these results may together support the claim:

$\mathcal{H}_{\text{f}}'$: x causes e.[39]

From this reasoning, increased variability in the replications will in principle, increase the scope of the hypotheses that would be supported by positive findings.

[39] This should not be understood as a strong metaphysical claim, and instead as a claim about reliable association.

5.4.1 Arguments from error, using close replications

If the primary hypothesis is passed *severely* over several different experiments that vary with respect to factors x, y, and z, the findings can analogously generate an argument from error to a stronger hypothesis. For example:

$\mathcal{H}_{\dagger}$:Subjects exposed to stimulus S have a higher rate of behavior B than those who are not [that is not attributable to x, y, z].

The construction of the above hypothesis serves to highlight the point that if the experiment finds consistent results despite a great deal of variation in its construction, this will support the inference that the hypothesized effect is not an artifact of those varied factors. In accordance with the severity requirement, an argument to having support for an effect could be constructed as follows:

1) Suppose that a set of *n* experiments with varying designs $\{d_1, d, \ldots d_n\}$ each pass $\mathcal{H}$ severely with e.
2) Further, suppose that $\{d_1, d_2, \ldots d_n\}$ vary such that each p_x controls for an error or source of bias $\{b_1, b_2, \ldots b_n\}$ that could account for e.
3) It is possible for any individual d_x to report an erroneous outcome.
4) Under the conditions in (1-2), there is a reduced probability of $\{d_1, d_2, \ldots d_n\}$ all (1) reporting mistaken results that (2) happened to all mistakenly report e, in a world where $\mathcal{H}$ is the case.
5) Under the conditions in (1-2), the replicability inquiry had a high probability of revealing an error (of accepting $\mathcal{H}$ when $\{b_1 \vee b_2 \vee \ldots \vee b_n\}$ is responsible for e) if it were present, but did not detect one nonetheless .
6) ES : $\mathcal{H}$ (and equivalent hypothesis $\mathcal{H}'$: *an error is absent*) is warranted by the data if and only to the extent that the scientific inquiry 1) had a high probability of revealing an error that is present and 2) did not detect such an error.[40]

C) There is strong evidence for $\mathcal{H}$.

This is an idealized case but it should get across the general structure of how one could use a number of probative replications to warrant inferences. The actual reasoning may vary, as researchers work with limited information from studies with varying levels of error control. For

---

[40] (D. G. Mayo 1994, 273)

example, a single close replication can be highly informative if it is able to introduce additional experimental constraints and replace problematic aspects of experimental design with more reasonable-seeming factors. The amount of replications performed, and the variability introduced, is good or bad *depending on the degree to which it allows one to make a strong argument from error.*

Consider a recent attempt to reproduce the results of a study on intelligence priming. Dijksterhuis and van Knippenberg (1998) had claimed to show that priming individuals with a "professor" stereotype increased performance on measures of intelligent behavior, whereas priming with the stereotype of "soccer hooligans" produced an opposite effect. The independent variable they used was performance on a general knowledge test – specifically, "a questionnaire with 42 difficult multiple-choice questions borrowed from the game Trivial Pursuit" (868).

A replication team questioned the legitimacy of this measurement – while memory and problem-solving seem to be affected by the allocation of mental resources, trivia appears to be the sort of thing a participant simply knows or doesn't know (Shanks et al. 2013, 2). They replaced this measurement with a critical thinking test, and found that they were unable to generate significant results, despite using a more powerful test than the original authors (*ibid.* 2, 7). Intuitively, this replication outcome is more informative than it would have been were the experiment a complete clone. The change in measure and increase in sample size meant the null hypothesis was subjected to a more severe test than in the original study; Shanks et al. seem to have constructed a more probative inquiry overall.

5.5 Conceptual replications

The divisions between different kinds of replication are ambiguous; replications start being referred to as *conceptual replications* when they become different enough from the original to be regarded as a different kind of experiment. Under error-statistical reasoning, they

would play an essential role in arguing for a hypothesis beyond the prediction that a treatment will have a hypothesized effect or that there is some correlation among a group of variables. The argument from error would a similar pattern to the one shown in section 4.2.1, except conceptual replications use a variety of different techniques to test the hypothesis of inference. Using different instruments and strategies, in addition to different samples and varied procedures designed to rule out errors particular to certain experiments, lowers the probability that a consistent result is artifactual. One may use this argument to severely probe a general research hypothesis such as:

$\mathcal{H}_{\mathrm{r}}$: Increasing self-esteem causes an increase in prosocial behavior.

This hypothesis is warranted through the support for its twin:

$\mathcal{H}_{\mathrm{r}}$ *: An error is not responsible for the results/is absent.

$\mathcal{H}_{\mathrm{r}}$ will be warranted to the extent that the replication inquiry was sufficiently probative. It can certainly be false, regardless of how impressive the agreement is between sub-inquiries. One can certainly imagine cases where the methods testing $\mathcal{H}_{\mathrm{r}}$ all fail to rule out a confounding factor, for example. It would still be the case that making such an argument would be desired for someone to accept a higher-level theoretical claim.

I do not mean this section to have been an exhaustive account of the role of replications in psychological research; my intention was to use the error-statistical framework to show how replication studies can be used in arguments from error to particular hypotheses. I will now show that framing replicability and reproducibility inquiries as arguments from error can help resolve a handful of significant problems brought up by the replication debates, through application of the informal severity requirement. For the same reasons that fallacies of rejection and fallacies of acceptance are problematic, certain interpretations of replication results will also be unwarranted.

I will begin the next section by examining the problem of interpreting negative results on the statistical level, and showing how to avoid committing the interpretive mistakes associated with them. I will then examine a challenge for the interpretation of negative results in the context of replication studies, namely the threat of dismissing a substantive hypothesis without warrant when presented with negative replication results. I argue that the apparent intractability of the problem is tied to a falsificationist position under which we cannot be warranted in making inferences from negative results. Worries about variation in close replications can be collapsed into a general Duhemian problem. I argue that the severity assessment performed to learn from non-significant results can act as a model for less formal assessments on the level of the research hypothesis, to allow for severe inferences about hypotheses in the face of conflicting replication results.[41] The rules for determining what hypothesis would have passed severely in a statistical context will have informal counterparts when it comes to replication arguments.

[41] See Mayo & Miller 2008.

6 Problems interpreting negative results

Non-significant results are ambiguous; as the aphorism goes, 'absence of evidence is not the same as evidence of absence': I may fail to find a statistical effect, but that does not mean there is necessarily warrant for accepting the hypothesis that the effect is not there. In the desire to avoid affirming the consequent – or to endorse a strong Popperian line – some methodologists have promoted the argument that because non-significant results are ambiguous, they cannot tell us about an effect being tested. Consider the following passage from Mark Leary's *Introduction to Behavioral Research Methods:*

> "You might think that results finding certain variables are not related to behavior – so-called null findings – would provide important information. After all, if we predict that certain psychological variables are related, but our data show that they are not, haven't we learned something important?
>
> The answer is no, for as we have seen, data may fail to support our research hypotheses for reasons that have nothing to do with the validity of a particular hypothesis. As a result, null findings are usually uninformative regarding the hypothesis being tested. Was the hypothesis disconfirmed, or did we simply design a lousy study? Because we can never know for certain, journals generally will not publish studies that fail to obtain effects." (2004, 24)

In the discussion about the replicability crisis, some have taken even stronger positions. Jason Mitchell, a neuroscientist interested in the replication debate, insists we should never publish null replication findings, because "null findings cannot distinguish between whether an effect does not exist or an experiment was poorly executed, and therefore have no meaningful evidentiary value even when specified in advance" (2014, "On the Emptiness of Failed Replications"). Though this view is extreme, it reflects a general attitude that the inability to perfectly rule out the influence of other variables precludes the possibility of learning anything at all.

On reflection, it may seem odd that a strict version of this position is so regularly endorsed. This is because there are clearly cases in which absence of evidence *is* evidence of absence, by any commonly understood use of the word. If I am trying to determine if there is a murderer hiding in my bedroom, and he isn't under my bed, in the closet, or any other plausible-seeming hiding place, I seem to have evidence that the murderer isn't actually there.[42]

The equivalent of the thorough room-check for a statistical significance test would be a test with high *power*. Prior to performing any experiment, one can use the effect size and significance criterion to determine the size of the sample necessary to detect an effect of a particular magnitude.[43] Such a test has a strong capacity for finding discrepancies from the null. If *T* has a very high probability of finding an effect when it is present, *and the null hypothesis still is not rejected*, we have at least some warrant for thinking that a discrepancy of that particular size is, in fact, lacking.

This warrant obviously will not transfer to all cases of null results. Severity reasoning requires *T* to have the capacity to detect discrepancies from *H* in order for passing with *e* to warrant the inference to *H*. One can begin to *prudently* claim evidence in favor of H_0 with framing the null as the assertion that a discrepancy is absent. Mayo and Cox have already outlined a principle in accordance with the severity requirement:

> FEV(ii): A non-significant p-value is evidence of the absence of a discrepancy δ from the null hypothesis only to the extent that there would have be a high probability of having found a smaller-p-value were δ to exist (Mayo & Cox 2006, 9-10).

---

[42] I would probably not concern myself with the outlandish possibility that he was an invisible murderer, or that he had, for the sake of my check, shrunk down to a microscopically tiny size.

[43] Tressoldi (2012) argues that given the estimated effect sizes of many psychological phenomena, a massive proportion of published studies – in some cases the great majority – lacked a sufficiently large sample size for the power needed to detect the relevant effects.

This provides a useful guide to whether there is support for the null or a discrepancy smaller than a certain amount. If the test has a high probability of finding a genuine divergence from the null hypothesis, but nonetheless fails to do so, we are warranted in inferring that a discrepancy beyond the established threshold is not present. If we are seriously motivated to avoid messing up, the error-statistical account tells us we should not be content to simply accept the null when it is fails to be rejected. To accept the null hypothesis when the test lacks the capability to detect a divergence from it (allowing it to detect an effect) is to commit what Mayo and Spanos refer to as a *fallacy of acceptance* (2006, 338).

Even if *T* has too low a power to find an effect, or the observed departure from the null is beneath the threshold for significance, a departure from the null may warrant ruling out some discrepancy γ from H_0. When the null is not rejected, we can ask, what is the largest γ that we would be warranted to infer with the same data *e?* In cases where it is very important to rule this out beyond a certain size, one must determine the severity with which the test might have passed the different hypothesis H^*:$(\mu_1 - \mu_2) = \delta$.[44] In other words, we calculate the probability of finding a result that is more extreme (a worse fit with H_0) than $d(x_{obs})$, if H^* were the case: Severity = $P(d(X) > d(x_{obs}); (\mu_1 - \mu_2) = \delta.)$

Now, I will show that similar reasoning that directs this process in the case of non-significant statistical results can be fruitful for understanding and addressing a seemingly un-related problem in the replication debate.

6.1 The problem of varying replications

When a direct replication does not reproduce the findings of an earlier study, the original researchers may choose to deny the experiment was correctly replicated in the first place. The

[44] (D. G. Mayo and Cox 2006, 16–17)

results of a replication, they argue, are uninformative if the procedure and general experimental design differ from the original experiment in some important aspect (which may be impossible to predict prior to data collection). As the Reproducibility Project notes, infinitely many experimental conditions could be necessary for bringing about an experimental outcome; "[t]he key question," they write, "is whether a failure to replicate could plausibly be attributed to any of these differences" (Aarts et al. 2013,18).

In one contested case, researchers were unable to reproduce results of a priming study that investigated whether increasing peoples' feelings of cleanliness would increase the severity of their moral judgments. Simone Schnall, the lead investigator, called attention to the fact that some of the replicators failed to control for interfering factors by performing the experiment online (2014). The result is that they could not ensure participants were focused on the priming task, and, as she further explains:

> "...the specific goal was to induce a sense of cleanliness. When participants do the study online they may be surrounded by mess, clutter and dirt, which would interfere with the cleanliness priming. In fact, we have previously shown that one way to induce disgust is to ask participants to sit at a dirty desk, where they were surrounded by rubbish, and this made their moral judgments more severe (Schnall, Haidt, Clore & Jordan, 2008). So a dirty environment while doing the online study would counteract the cleanliness induction." (Schnall, *"Further Thoughts on Replications, Ceiling Effects and Bullying"*)

This seems like an understandable concern. But in other cases, it is hard to tell whether the alterations would ever be relevant to the study outcome.[45] As Stanley Klein (2014) notes, this is especially a problem in social priming, where it is assumed that prima facie irrelevant factors like the kind of instrument used to time participants will determine whether a replication finds positive results (328–331). Consider the reaction that Ferguson et al. gave when the Many Labs

[45] For examples, see John Bargh's archived 2012 essays from *Psychology Today, "Angry Birds"* and *"Nothing in their Heads."* These were academic blog posts; much of this debate has unfolded through such nontraditional mediums.

Replications Study could not reproduce their results. In the original study, it had been hypothesized that if individuals were exposed to a flag prime, they would experience a political shift specifically towards Republicanism, as measured by scores on a political questionnaire.[46] When replicators could not find the same effect, the original authors responded that this was not informative. This is because according to the hypothesis,

> [Flags] activate knowledge related to one's nation that is shaped by the prevailing political atmosphere, which is hardly inert. The original experiment was run in 2009 – shortly after the first African-American was sworn in as president of the US – whereas the ManyLabs was run 4 years later, in the 5th year of his presidency. (Ferguson, Carter, and Hassin 2014, 301)

The researchers suggested that changes to the political climate altered the effects of the priming conditions, so that the direct replication was merely conceptual. They also pointed out that the Many Labs studies had subjects participate in a series of experiments, some of which mentioned the United States. This meant any participants who did not complete this study first could have been pre-primed; and the sample size, when reduced to just that set, would lack the power to test the theoretical model that the authors thought would make the effect significant (2014, 302).[47]

6.2 A Duhem-Quine problem for replication?

Such responses reflect genuine challenges to making inferences from experimental results. In an upcoming paper, Earp and Trafimow (2015) formalize the problem in terms of the

---

[46] Klein et al. only replicated the second experiment of the original study. From the first experiment, the original authors had concluded that brief exposure to a tiny American flag was able to cause long-term shifts in political beliefs towards Republicanism. Study 2 was designed to rule out the possibility that the participant attitudes had merely shifted towards whoever was currently in power (Ferguson, Carter & Hassin 2011).

[47] This was an unusual response considering in the original study, this effect was detectable for up to *eight months* following exposure to the flag. It is not clear why the threat of prior "US" priming would be such a problematic confounding variable for the replication but not for the original experiment, which took place during election season (Carter, Ferguson, and Hassin 2011, 1015–1016).

Duhem-Quine thesis (15) . This problem poses a difficulty for scientists when they try to correctly assign responsibility for an outcome that goes against their hypothesis.

The problem exists on both the methodological and logical level, as it reflects how *modus tollens* operates for any antecedent that is a conjunction. Consider the following argument structure, where H_R is the primary hypothesis, $A_1...A_n$ are auxiliary hypotheses, C_p is a ceteris paribus clause, and O is an experimental outcome:

1) $\left(H_R \wedge \left(A_1, A_2, \ldots A_n\right) \wedge C_p\right) \supset O$

2) $\sim O$

C) $\sim\left(H_R \wedge \left(A_1, A_2, \ldots A_n\right) \wedge C_p\right)$

From the above we see that the observation ~O does not falsify H_R. Instead, it falsifies the conjunction of H_R and a whole slew of auxiliary assumptions. These assumptions include everything from claims about experimental conditions to theories associated with measurement instruments. Because the entire conjunction is subject to confirmation or disconfirmation, H_R could always survive refutation in the face of negative results by having blame assigned to the auxiliaries or rejecting the ceteris paribus clause.[48] The problem is even further complicated when hypotheses involve statistical claims, and ~O isn't actually logically inconsistent with the conjunction.

Psychological researchers are well aware of this sort of problem.[49] In fact, they use it to justify the file-drawer problem, as seen in the earlier passage from Leary (2004). However, there

[48] See Søberg 2005, p. 3, and Harding 1976, p. ix–xi.

[49] Psychologists do not always refer to the Duhem-Quine thesis by name, but the replication literature is rife with

seems to be a perception that the underlying reasoning takes on a unique character in replicability checks, which may explain the hard emphasis on direct replication in the current debate. According to Earp and Trafimow, in the case of interpreting the results of direct replications, *ceteris paribus* would state that there are no changes between the original and the replication that would contribute to negative results (2015, 17).

In the case of direct replication, one tries to falsify the claim that a particular experiment, with a specific procedure and specific set of essential conditions, will yield a particular outcome. To try and model this, we'll use the symbol A_O for the set $A_1, A_2, \ldots A_n$ above. We also include C_r, as a ceteris paribus clause:

1. $(H_R \wedge A_O \wedge C_r) \supset O$

2. $\sim O$

---

3. $\sim(H_R \wedge A_O \wedge C_r)$

The replicability check tests a conjunction of the research hypothesis with all of the auxiliary assumptions of the original experiment (including the ceteris paribus clause), plus a new assumption – that there are no important differences between the replication study and the original that would prevent *O* from obtaining. If this is the case, in order to have confidence in the findings of a replication study, its auxiliary assumptions must be close enough to the original experiment's assumptions that it would be hard to plausibly blame negative findings on a difference between the two studies (17). A similar attitude also appears in LeBel and Peters (2011), who agree that "confidence in a negative result increases directly the closer the design of the replication study is to that of the original study" (376).

---

discussion of Duhemian problems. For a recent direct reference and discussion, see (LeBel and Peters 2011) in the context of interpreting results from parapsychology.

The changes made in the replication study will lead to auxiliary assumptions that are of varying reasonableness. For example, I might hand over a task handled by humans in the original study to a program in the replication, with the thought that a machine will be more consistent and reliable than a graduate assistant. I would be removing the assumption that the grad student is making an accurate observation, and introducing the assumption that a machine programmed to perform this exact test, that I recently calibrated, will give me accurate readings. This seems like a likely improvement over the original design – the problem, Earp and Trafimow argue, is that it is not always clear how to judge which auxiliary assumptions will be 'better,' or when they are importantly relevant at all. This means that when a researcher performs a direct replication and their data analysis yields a different result than the original researcher, she will be unable to take this as evidence against the research hypothesis.

6.2.1 Earp and Trafimow's 'demonstration' of how we learn from replications

This is at tension with the intuition that even if individual studies cannot be definitive, enough negative results should provide evidence against the findings of the original study (E&T 17-18). Earp and Trafimow take this as a strike against falsificationist accounts of experimental learning, and propose an alternative Bayesian framework to reconstruct how negative replication results could decrease confidence in past findings (moderated by the quality of auxiliary conditions). To illustrate, imagine two psychological researchers, Anne and Beth. Beth decides to perform a close replication of one of Anne's studies but is not able to reproduce her results. If Anne's colleague Catherine admires Anne's work, she might still have confidence in findings, even after finding out about Beth's negative results. This is because of the high prior probability she puts on Anne's findings being correct.

This confidence level corresponds to the posterior probability of the original findings being true given that Beth could not replicate the study: P(Findings|~Replicate) (18). One could

alternatively determine the posterior confidence ratio that the findings are true versus not true, given the inconsistent results and the researcher's prior confidence ratio. Either way, as the number of replication failures increases, a rational researcher will have less and less confidence in the original findings. Catherine will eventually reach a point where she is more confident that there is no effect (or that the original hypothesis is otherwise false) than that the original hypothesis was true.

Earp and Trafimow's account requires us to make two assumptions. The first is that it is considerably more likely for replications to fail if the original findings were false than if they were true:

$$\frac{P(\text{failed replication}|\text{original findings were true})}{P(\text{failed replication}|\text{original findings were false})} \ll 1$$

The next is that, as a close replication study's auxiliary assumptions improve towards an ideal case for the ceteris paribus clause, the confidence ratio $\frac{P(\text{failed replication}|\text{findings true})}{P(\text{failed replication}|\text{findings false})}$ deviates further from 1. It follows that Beth's failed replication will lower Catherine's confidence in the truth of Anne's results, and that for any replications, the degree to which Catherine's confidence is lowered by a failed replication will be dependent on the quality of the auxiliary assumptions of that study.

Under these additional assumptions, negative findings will cause a researcher to lose confidence in the accuracy of the findings (and the hypothesis of a reliable effect), and the degree to which this will be the case will depend on the strength of the auxiliary assumptions (19). This is supposed to prove that replications are informative dependent on the quality of the studies in a way that allows us to isolate problematic auxiliary assumptions over time. The problem is that Earp and Trafimow's solution does not actually show this. It may serve as a model for how researchers make decisions about hypotheses, separate from how warranted those

decisions ultimately are, but there are dangers with justifying the learning process with reference to how individual scientists feel about particular replications. Though in their model, even stubborn individuals will shift their positions over time, it is not clear that this is so in real life; even so, in many cases warrant for rejecting a finding as genuine would fail to develop over any reasonable number of replication studies, making blame allocation for negative results an extremely drawn-out process. Further, the posterior probability is strongly dependent on the prior probability that gets assigned – the whole enterprise is subjective.

Finally, a great deal of their position does not make sense if we are merely interested in increasing our confidence in individual, strictly-defined studies. Why should a change in auxiliary assumptions increase our confidence in replication results if they are just meant as an appraisal of a previous experiment?[50] Their insight about the importance of different auxiliary hypotheses can be easily incorporated into the error-statistical account as reflecting the importance of informal error control. In a passage quoted earlier, LeBel and Peters claimed that close replications were preferable because 1) researchers try to avoid changing aspects of the methodology between the studies, and 2) the differences that do exist between the studies (should) have been intentionally introduced. This helps narrow down where blame is assigned when a replication study fails, making direct replications "the most diagnostic test of whether an observed effect is real" (376). If arguments to the existence of reliable effects are arguments from error, then we do not have to resort to subjective confidence measures to explain why we would probably consider a great number of non-significant findings to be evidence against the

[50] If we are merely interested in scrutinizing the first set of results, then there are of course ways that a legitimate replication check might cast doubt on past findings. An example for "successful replications" would be when the effect size found by the replication study was small enough to render it highly unlikely for the original study to have picked up on it.

hypothesis supported by a single original study. It would be improbable for a set of powerful experiments to pass the null hypothesis if the research hypothesis were correct, especially if this support were consistent across variations in study design and experimental conditions, and the individual inquiries were highly probative (ruling out the influence of particular auxiliary conditions with high severity). At the same time, this understanding can accommodate the fact that a high replication rate may not warrant the inference to a real effect; the process must probe for error across levels of inquiry, because both poorly designed experiments and factors such as publication bias can create the appearance of a well-supported hypothesis.

6.3 How do we learn from non-significant replication results?

The error-statistical account rejects both the naïve Popperian understanding common to psychologists, and credentist explanations such as Earp and Trafimows'. The question of whether a negative replication result warrants us to accept that the research *hypothesis is not the case*, is a question of whether the hypothesis of its falsity passes a severe test with the evidence. Just as whether accordance with a null statistical hypothesis is supported by the data is largely dependent on how powerful the test was, a negative replication result is only evidence against the research hypothesis if, and to the extent that, the replication inquiry had a high probability of finding results consistent with that hypothesis were it true and found no such results.

6.3.1 Avoiding fallacies of acceptance for negative replications

One can apply reasoning that is structurally similar to the kind used to interpret non-significant statistical results at the level of research hypotheses under investigation in replication experiments. Suppose I am replicating an experiment that originally found a large effect size supporting hypothesis H^*. In my replication study, the statistical null H_0 is not rejected, and further, my results severely pass the hypothesis that there is no discrepancy from H_0 larger than

size we take to be meaningful. Do my results indict the original study? Are they even relevant to the same hypothesis?

To be consistent with requirements for severity, successfully interpreting a negative replication will require checking to see if the assumptions of the test were met. After this, the severity analysis for determining what hypothesis is warranted by the replication data will require comparison among the different sets of results.

6.3.2 What hypotheses do pass severely?

What hypotheses do pass severely? Recall the argument structure presented in the section on close replications. That was an ideal case, in which all of the individual experiments had passed *H*, with high severity. In a more realistic scenario, the evidence will likely be quite mixed, and the replication studies themselves will vary in quality. A negative result may be a poor indicator that *H* is false; but when *H* is unsupported by experiments that control for a particular factor, or experiments in a particular cultural setting, we might reasonably take this to warrant the inference that *H* is not the case. Just as in the case of failing to reject the statistical null hypothesis, we are directed to 1) consider whether or not the inquiry was powerful enough to detect such an effect, and 2) explore what hypotheses *would* have passed the replicability check severely.

Acknowledging a non-replication does not mean acknowledging that all hypotheses close to *H* are not warranted. It may be the case that a similar hypothesis to *H* passes severely, but with more limited scope. If this is the case, it is still experimental knowledge: we can still learn, with high severity, if we are warranted in making another inference.

6.4 What do we learn from positive replication results?

The same way that negative results may fail to warrant the inference that an effect is absent, researchers are also liable to mistakenly infer support for a substantive claim from the

consistent replication of experimental results. Earler I wrote that the error-statistical account instructs the researcher to break down large questions into smaller, more manageable questions with hypotheses amenable to severe probing. It is important to keep these separate because the same sort of results that severely pass local hypotheses will not necessarily pass substantive hypotheses.

6.4.1 What do we learn from large-scale replication projects?Avoiding fallacies of rejection

Consider Daniel Kahneman's suggestions for exonerating psychology from the replication crisis, where he recommends the use of a *daisy chain of replications*. He proposes that a group of labs comes together as a replication group in which each selects an experiment to be replicated and performs a replication experiment for the next lab in the chain (Kahneman 2012). Each replication should involve consultation with the original researcher to ensure the procedure is run correctly, along with supervision by them; that the sample sizes should be large enough to have a high-powered test; and that the study researchers should be respected, experienced members of the field. However, he does not say anything about checking the study design for appropriateness, adequate control and conditions, or similar considerations.

"Success (say, replication of four of the five positive priming results) would immediately rehabilitate the field," Kahneman writes (*ibid.*). This claim is somewhat ambiguous – he may only mean that *credibility* would be restored by successful replication attempts, not that the scientific status of these inquiries would be confirmed. However, researchers should be careful to avoid conflating a demonstration of replicability with a demonstration of realness. The suggested protocol, and the ones used in the reproducibility projects that have since been developed, only have the ability to warrant the sort of limited inferences described in section **Error! Reference ource not found.**.

If there is something wrong with an experiment's protocol such that it guarantees the generation of a positive result regardless of the truth of the research hypothesis, the results will be highly replicable – but only in a totally vacuous sense. As a result, even under the condition that one's analysis passes the statistical hypothesis severely, the inference to a substantive effect would not be warranted. Further, with the exception of a few cases, even if the individual inquiries severely probe the local research hypothesis, passing it severely will not be sufficient for inferring the relevant theoretical hypothesis. We cannot warrant the inference to a substantive hypothesis using direct replication *per se*; to do so requires the sort of argument that conceptual replications lend themselves to, or at the very least some additional error-probing inquiries.

7 Final thoughts

In this essay, I have tried to show that applying an error-statistical framework can help resolve issues related to the replicability crisis. In the course of this, I have also argued that replication inquiries support hypotheses via arguments from error.

One might be concerned that it is not clear why one would want to use this account to frame the problem, as opposed to Earp and Trafimow's Bayesian explanation, or the Popperian account that seems to be favored by many scientists. Here I would repeat my criticisms from 5.2.1, and emphasize that neither of these can currently offer recommendations for researchers.

The error-statistical account is also uniquely equipped to consider the sort of inference tools already used in experimental psychology – specifically null-hypothesis significance testing, Neyman-Pearson testing and confidence interval use. In the course of this paper I have avoided going into the details of calculating severity for several kinds of cases, but guides to doing so are readily available; see Mayo and Spanos (2006) for examples.

One might also worry that the severity requirement is overly harsh – that it is impossible to ever warrant a psychological research hypothesis, because it is impossible to ever fully rule out error. This objection is misguided. There is no intent here to claim that it is possible to rule out all possible sources of error from an experimental inquiry. In fact, it is regularly impossible to rule out potentially devastating ones, at least when moving from 'local' or descriptive hypotheses to inferences about theories. This does not mean all our inferences are unwarranted.

Imagine that award-winning experimentalist Max Diligence is interested in whether performance on a particular cognitive task, Task X, differs by gender. He sets out the research hypothesis:

H: Women score higher than men on Task X.

Now suppose Max goes out of his way to check that every possible aspect of the experiment is well-controlled. He checks the measurement scale for reliability. He goes out of his way to prevent the influence of known biasing factors, such as stereotype threat, that could block the effect from appearing. In short, he does everything he needs to do to confidently link a potential rejection of the statistical null hypothesis to support for his local research hypothesis.

Now suppose further that psychological laboratories are haunted by sexist ghosts. These ghosts conspire to make men perform worse on Task X than their female counterparts, by distracting them in ways that are not consciously perceptible. Assume they do this every time Max performs the experiment, resulting in a finding that is highly reproducible.

There are two points I would like to make here. The first is that the inference to the local hypothesis *H* is in no way undermined by the fact that ghosts are causally responsible; error probabilities are not dependent on substantive factors. The men are genuinely performing worse on the task than the women are.

My second, probably more contentious point is that the fact that an unknown factor could render the further inference to a particular theory false *does not necessarily mean we should consider that inference unwarranted.* Our motivation to find out things about the world – including information about experimental effects – does important work here: it would surely be strange to say that the possibility of confounding factors makes it not worthwhile to try to improve error control and account for sources of bias.

The severity requirement provides a guide to whether we are warranted in accepting an inference, and sometimes directly indicates how another, more qualified inference might be warranted. It is true that it is impossible to rule out all possible sources of error – but part of the point of experimental design is constraining the practical threats that must be dealt with. We

must be able to expect a basic level of self-criticism on the behalf of psychological researchers, especially those who set out to critically scrutinize past experimental findings.

References

Aarts, Alexander A, Anita Alexander, Peter Attridge, Elizabeth Bartmess, San Francisco, Frank A Bosco, Benjamin Brown, et al. 2013. “The Reproducibility Project: A Model of Large-Scale Collaboration for Empirical Research on Reproducibility.” 1–36.

Alexander, Anita, Michael Barnett-Cowan, Elizabeth Bartmess, Frank A Bosco, Mark J Brandt, Joshua Carp, and Jesse J Chandler. 2012. "An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science." *Open Science Collaboration.*

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. 2008. “Reporting Standards for Research in Psychology: Why Do We Need Them? What Might They Be?” *The American Psychologist* 63 (9): 839–51.

Bem, Daryl J. 2011. “Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect.” *Journal of Personality and Social Psychology* 100 (3): 407–25.

Berger, James O, and RL Wolpert. 1988. *The Likelihood Principle: A Review, Generalizations, and Statistical Implications.* Lecture Notes - Monograph Series. Hayward, California.

Brandt, Mark J., Hans IJzerman, Ap Dijksterhuis, Frank J. Farach, Jason Geller, Roger Giner-Sorolla, James a. Grange, Marco Perugini, Jeffrey R. Spies, and Anna van ’t Veer. 2014. “The Replication Recipe: What Makes for a Convincing Replication?” *Journal of Experimental Social Psychology* 50: 217–24.

Carter, Travis J, Melissa J Ferguson, and Ran R Hassin. 2011. “A Single Exposure to the American Flag Shifts Support toward Republicanism up to 8 Months Later.” *Psychological Science* 22 (8): 1011–18.

Cartwright, N. 1991. “Replicability, Reproducibility, and Robustness: Comments on Harry Collins.” *History of Political Economy* 23 (1): 143–55.

Cohen, Jacob. 1994. “The Earth Is Round ($p < .05$).” *American Psychologist* 49 (12): 997–1003.

Collins, Harry M. 1984. “When Do Scientists Prefer to Vary Their Experiments?” *Studies in History and Philosophy of Science* 15 (2): 169–74.

Dijksterhuis, Ap, and Ad van Knippenberg. 1998. “The Relation between Perception and Behavior, or How to Win a Game of Trivial Pursuit.” *Journal of Personality and Social Psychology* 74 (4): 865–77.

Earp, Brian D, and David Trafimow. 2015. “Replication, Falsification, and the Crisis of Confidence in Social Psychology.” Accessed October 1, 2014. https://www.academia.edu/10078878/Replication_falsification_and_the_crisis_of_confidence_in_social_psychology.

Ferguson, Melissa J., Travis J. Carter, and Ran R. Hassin. 2014. “Commentary on the Attempt to Replicate the Effect of the American Flag on Increased Republican Attitudes.” *Social Psychology* 45 (4): 299–311.

Filliben, James J., ed. 2003. “Critical Values of the Student’s T Distribution.” In *E-Handbook of Statistical Methods.* SEMATECH. Accessed January 6, 2015. http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm.

Fisher, Ronald A. 1926. “The Arrangement of Field Experiments.” Journal of the Ministry of Agriculture of Great Britain 33: 503–13.

Francis, Gregory. 2012a. “The Psychology of Replication and Replication in Psychology.” *Perspectives on Psychological Science* 7 (6): 585–94.

Francis, Gregory. 2012b. “Publication Bias and the Failure of Replication in Experimental Psychology.” *Psychonomic Bulletin & Review* 19 (6): 975–91.

Galak, Jeff, Robyn a Leboeuf, Leif D Nelson, and Joseph P Simmons. 2012. “Correcting the Past: Failures to Replicate Ψ.” *Journal of Personality and Social Psychology* 103 (6): 933–48.

Gelman, Andrew, and Eric Loken. 2014. “The Statistical Crisis in Science: Data-Dependent Analysis—a ‘Garden of Forking Paths’— Explains Why Many Statistically Significant Comparisons Don’t Hold Up.” American Scientist 102 (6): 2–6.

Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science.* Cambridge: Cambridge University Press.

Harding, Sandra G. 1976. *Can Theories Be Refuted? Essays on the Duhem-Quine Thesis*. Edited by Sandra G Harding. Dordrecht, Holland: D. Reidel Publishing Company.

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams, Štěpán Bahník, Michael J. Bernstein, Konrad Bocian, et al. 2014. “Investigating Variation in Replicability.” *Social Psychology* 45 (3): 142–52.

Klein, Stanley B. 2014. “What Can Recent Replication Failures Tell Us about the Theoretical Commitments of Psychology?” *Theory & Psychology* 24 (3): 326–38.

Lambdin, Charles. 2012. “Significance Tests as Sorcery: Science Is Empirical - Significance Tests Are Not.” *Theory & Psychology* 22 (1): 67–90.

Leary, Mark R. 2004. *Introduction to Behavioral Research Methods*. 4th ed. Boston: Pearson Education, Inc.

LeBel, Etienne P., and Kurt R. Peters. 2011. “Fearing the Future of Empirical Psychology: Bem’s (2011) Evidence of Psi as a Case Study of Deficiencies in Modal Research Practice.” *Review of General Psychology* 15 (4): 371–79.

Makel, Matthew C., Jonathan A. Plucker, and Boyd Hegarty. 2012. “Replications in Psychology Research: How Often Do They Really Occur?” *Perspectives on Psychological Science* 7 (6): 537–42.

Mayo, Deborah G. 1991. “Novel Evidence and Severe Tests.” *Philosophy of Science* 58 (4): 523–52.

Mayo, Deborah G. 1994. “The New Experimentalism, Topical Hypotheses, and Learning from Error.” *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1: 270–79.

Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge.* 1st ed. Chicago: University Of Chicago Press.

Mayo, Deborah G, and David R Cox. 2006. “Frequentist Statistics as a Theory of Inductive Inference.” *IMS Lecture Notes - Monograph Series*, 1–28.

Mayo, Deborah G, and Aris Spanos. 2006. "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *The British Journal for the Philosophy of Science* 57 (2): 323–57.

Mayo, Deborah G, and Aris Spanos. 2011. "Error Statistics." In *Handbook of the Philosophy of Science, Volume 7: Philosophy of Statistics.*, edited by Prasanta S Bandyopadhyay, Malcolm R Forster, Dov M Gabbay, Paul Thagard, and John Woods, 7:153–98. Elsevier B.V.

Mayo, Deborah G, and Jean Miller. 2008. "The Error Statistical Philosopher as Normative Naturalist." *Synthese* 163: 305–14. doi:10.1007/s.

Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46 (September 1976): 806–34.

Mitchell, Jason. 2014. "On the Emptiness of Failed Replications." Accessed January 2, 2015, http://wjh.harvard.edu/~jmitchel/writing/failed_science.htm

Nosek, B. a., J. R. Spies, and M. Motyl. 2012. "Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability." *Perspectives on Psychological Science* 7 (6): 615–31.

Pashler, H., and C. R. Harris. 2012. "Is the Replicability Crisis Overblown? Three Arguments Examined." *Perspectives on Psychological Science* 7 (6): 531–36.

Popper, Karl. 1953. "Science: Conjectures and Refutations." *In Conjectures and Refutations: The Growth of Scientific Knowledge*, 33–65. New York: Basic Books.

Popper, Karl. 2005. *The Logic of Scientific Discovery.* Taylor & Francis E-Library. e-Book. London/New York: Routledge Classics.

Psillos, Stathis. 2008. "Cartwright's Realist Toil: From Entities to Capacities." In *Cartwright's Philosophy of Science*, edited by Stephan Hartman, Carl Hoefer, and Luc Bovens, 167–94. Routledge.

Ritchie, Stuart J, Richard Wiseman, and Christopher C French. 2012. "Replication, Replication, Replication." *The Psychologist* 25 (5): 346–49.

Schnall, Simone. 2014. "Further Thoughts on Replications, Ceiling Effects and Bullying." *University of Cambridge Department of Psychology Blog*. Accessed July 5, 2014. http://www.psychol.cam.ac.uk/cece/blog

Shanks, David R, Ben R Newell, Eun Hee Lee, Divya Balakrishnan, Lisa Ekelund, Zarus Cenac, Fragkiski Kavvadia, and Christopher Moore. 2013. "Priming Intelligent Behavior: An Elusive Phenomenon." *PloS One* 8 (4): e56515.

Simmons, Joseph P, Leif D Nelson, and Uri Simonsohn. 2011. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22 (11): 1359–66.

Simonsohn, Uri. 2014. "Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also." *SSRN Electronic Journal*, 1–10. doi:10.2139/ssrn.2374040.

Simonsohn, Uri, Leif D Nelson, and Joseph P Simmons. 2013. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534–47.

Søberg, Morten. 2005. “The Duhem-Quine Thesis and Experimental Economics: A Reinterpretation.” *Journal of Economic Methodology* 12 (329): 581–97.

Tressoldi, Patrizio E. 2012. “Replication Unreliability in Psychology: Elusive Phenomena or ‘Elusive’ Statistical Power?” *Frontiers in Psychology* 3 (January): 1–5.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L J van der Maas. 2011. “Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi: Comment on Bem (2011).” *Journal of Personality and Social Psychology* 100 (3): 426–32.

Appendix A

The following is a simple example calculating severity after H_0 is rejected. Imagine we are interested in checking to see if the treatment group n_1 scores higher than the control group n_2 on a particular measure. If both n_1 and $n_2 = 50$, and $\alpha = .05$, then for a one-tailed independent means t-test the rejection rule would be:

$$\text{Reject } H_0\text{: } (\mu_1 = \mu_2) \text{ whenever } T\alpha > 1.661.$$

Say I collect my data and find that $\bar{X}_{\text{stimulus}} = 30$ and $\bar{X}_{\text{control}} = 27$. If my pooled standard error of difference is 6, then my t-score is:

$$t = \frac{[30 - 27]}{6\left(\sqrt{1/25}\right)} = \frac{3}{1.2} = 2.5$$

H_0 is rejected at $p<.05$, since $2.5 > 1.661$. This warrants the hypothesis of a positive discrepancy from the null, H': $(\mu_1 - \mu_2) > \delta$.

At $t = 2.5$, $p \approx .007$, so H' will pass with a severity of .993. The severity with which a *particular* divergence δ would be warranted will decrease as the size of δ increases (as illustrated in section 3.2).