Date: 20240818, forthcoming with *Synthese* (ISSN 1573-0964)

Cohen's Convention, the Seriousness of Errors, and the Body of Knowledge in Behavioral Science

Aran Arslan[1] and Frank Zenker[2]

[1] Department of Philosophy, Boğaziçi University, 34342, Bebek, Istanbul, Türkiye
[2] College of Philosophy, Nankai University, Tianjin, P.R. China
Aran Arslan https://orcid.org/0000-0002-3014-6532
Frank Zenker https://orcid.org/0000-0001-7173-7964
Correspondence: Frank Zenker, frank.zenker@nanakai.edu.cn; fzenker@gmail.com

Abstract: An often-cited convention for discovery-oriented behavioral science research states that the general relative seriousness of the antecedently accepted false positive error rate of $\alpha = .05$ be mirrored by a false negative error rate of $\beta = .20$. In 1965, Jacob Cohen proposed this convention to decrease a β-error typically in vast excess of .20. Thereby, we argue, Cohen (unintentionally) contributed to the wide acceptance of *strongly uneven* error rates in behavioral science. Although Cohen's convention can appear epistemically reasonable for an individual researcher, the comparatively low probability that published effect size estimates are replicable renders his convention unreasonable for an entire scientific field. Appreciating Cohen's convention helps to understand why even error rates ($\alpha = \beta$) are "non-conventional" in behavioral science today, and why Cohen's explanatory reason for $\beta = .20$—that resource restrictions keep from collecting larger samples—can easily be mistaken for the justificatory reason it is not.

Keywords: false positive and false negative test results; inductive risk; null hypothesis significance testing; type I and type II error; utility; value-free science

1. Introduction

As a review of publications in behavioral science would show, a widely cited (if less thoroughly implemented) convention on the error rates in discovery-oriented research originates with the statistician Jacob Cohen. In his book *Statistical Power Analysis for the Behavioral Sciences* (Cohen, 1969), this convention is supported by

what Cohen (1965) called the *general relative seriousness* of false positive and false negative errors (aka α- and β-error rates or errors of Type I and II).

> "It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value [$(1 - \beta)$ =] .80 be used. This means that β is set at .20. […] This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-9). The chief among them takes into consideration the implicit convention for α of .05. The β of .20 is chosen with the idea that the *general relative seriousness* of these two kinds of errors is of the order of .20 / .05, i.e., that Type I errors are of the order of four times as serious as Type II errors." (Cohen, [1]1969, 1988, p. 56; *italics added*)

As in Cohen (1965), the general relative seriousness of an error would also later be understood in terms of its *cost*:

> "The author has proposed a convention for desired power of .80 (Cohen, 1965, 1969). It is suggested for use when no other value is suggested by the ad hoc demands of the research, and for methodological surveys and the like. Taken together with the α = .05 convention, it suggests the stance that Type I errors are about four times as "*costly*" as Type II errors, i.e., β / α = .20 / .05 = 4." (Cohen, 1970, p. 825; *italics added*)

Later yet, in a five-page review wherein "the sample sizes necessary for .80 power to detect effects [of various size] […] are tabled for 8 standard statistical tests" (Cohen, 1992, p. 155), a review that amasses some 60,000 citations today (Google Scholar), the term 'costly' is connected (more explicitly than in Cohen's earlier writings) to the cost of collecting a sample. Cohen holds that, while "a materially smaller value than [1 – β =] .80 would incur too great a risk of a Type II error," a "materially larger value would result in a demand for [the sample size] *n* that is likely to exceed the investigator's resources" (Cohen, 1992, p. 156). Part of the motivation for Cohen's convention, then, is the statistical fact that, other things equal, securing a small(er) β-error rate requires a large(r) sample.

Here we evaluate Cohen's convention within the debate on the value-laden character of science. We argue that the collectively unreasonable consequence of Cohen's convention is that, while *discoveries*—understood as independently reproducible effect size estimates of sufficient size—presuppose *small* and *even* error rates ($\alpha = \beta << .05$), Cohen's convention (unintentionally) led to the wide acceptability in behavioral science of *strongly uneven* error rates. Although the justification Cohen offers can appear epistemically reasonable for individual researchers, the convention is collectively unreasonable for an entire scientific field because it entails a low probability that published effect size estimates are replicable.

Consistent with how others explain the replication crisis in behavioral science, appreciating the role of Cohen's convention not only helps to understand why even error rates ($\alpha = \beta$) today are "non-conventional" in behavioral science but also why Cohen's explanatory reason for $\beta = .20$—that resource restrictions often keep from collecting larger samples—is easily mistaken for the justificatory reason it is not.

2. Discoveries, Cohen's convention, and the body of scientific knowledge

2.1 Discovery-oriented hypothesis testing research

Because raw measurement scores ("observations") are subject to error, the measurement scores of behavioral responses (as sampled from a population) must be related to an empirical hypothesis via intermediate statistical inference procedures. As these procedures transform raw measurement scores into probability density distributions ("data"), it is revealed as a "naïve fantasy that data have an immediate relation to phenomena of the world […], that they are the facts of the world directly speaking to us […]" (Longino, 2020, p. 391). Instead, when data inform hypothesis-related decisions, the "right measure of evidential support generally has a probabilistic character" (Diez, 2011, p. 105; see Krüger et al., 1987).

Despite the increasing prominence in behavioral science of the Bayesian approach to statistical inference (Fienberg, 2016), the default statistical paradigm is null hypothesis significance testing (NHST) (Gigerenzer, 1987; 2004; Morrison & Henkel, 1970). In Fisher's (1956) version of NHST, data are compared narrowly to a null hypothesis (H_0) that normally states a zero effect/correlation between variables. Whereas in the Neyman-Pearson version (Neyman & Pearson, 1967), which advances Fisher's (1956), data are additionally compared to an alternative

hypothesis (H_1), stating an effect/correlation that is non-zero (directional H_1) or of some definite strength (point H_1).

A crucial limitation is that both versions of NHST only inform the decision to *maintain* or *reject* H_x (x = 0, 1), given the rejection criterion that the probability of data (D) in view of H_x is smaller than the statistical significance level. In the case of the conventional statistical significance level p = .05, for instance, '$p(D, H_x) < .05$' means: the probability of D in view of H_x is smaller than .05. Beyond this limitation lies the decision to *accept* H_x as probabilistically supported by data. It generally requires defining a support threshold on such Bayesian measures as the likelihood ratio, $LR_{H1/H0} = [L(H_1 \mid D) / L(H_0 \mid D)] = [P(D, H_1) \times P(H_1) / P(D, H_0) \times P(H_0)]$, or the Bayes factor, $BF_{H1/H0} = [P(D, H_1) / P(D, H_0)]$ (Edwards, Lindman & Savage, 1963; Edwards, 1972; Witte & Zenker, 2017; see our footnotes 1 and 2).

While the decision to accept H_x is what NHST cannot provide, the decision to maintain H_0 entails the decision to reject H_1, and *vice versa*. Both kinds of decisions are associated with two types of error: a true hypothesis may be rejected and a false hypothesis maintained. The long-run chances of rejecting a true H_0 is what the Neyman-Pearson version of NHST calls the α-error rate and those of rejecting a true H_1 the β-error rate. A hypothesis test thus has four possible outcomes (Table 1).

Table 1: Confusion matrix of the possible outcomes of an NHST hypothesis test

	H_0 is *maintained*	H_0 is *rejected*
H_0 is *true*	correct decision or test result	α-error false positive error Type I error
H_0 is *false*	β-error false negative error Type II error	correct decision or test result

2.2 Cohen's Convention

Forwarded against the background of the Neyman-Pearson version of NHST, the *comparative* version of Cohen's convention states:

(1) The consequences of a false positive error, i.e., the mistaken rejection of a true H_0 hypothesis (α-error), are *more serious* than the consequences of a false negative error, i.e., the mistaken rejection of a true H_1 hypothesis (β-error).

Consequently:

(2) When statistically significant test results are reported, the ratio of the long-run chances of committing α- and β-errors can, in the absence of other considerations, be set *asymmetrically* in favor of minimizing the α-error rate.

The more informative, quantified version of Cohen's convention states that the probability of mistakenly rejecting a true H_0 (α-error) be set to *one-fourth* of that of mistakenly maintaining a false H_0 (β-error). Given the conventional $\alpha = .05$, this means accepting $\beta = .20$, i.e., a ratio of $\alpha / \beta = (.05 / .20) = 1 / 4$ (Fig. 1). Both values happen to be typical default settings of automated sample size planners (e.g., Kovacs et al., 2022).

According to Cohen, (2) follows from (1) because "the notion that failure to find something is *less serious* than finding something that is not there accords with the conventional scientific view" (Cohen, 1977; 1988, p. 56, *italics added*). When comparing these two undesirable events, failing to find what is there amounts to a *missed discovery*, and "finding" what is not there to a *mistaken discovery*. Once published in the scientific literature—here called the 'body of scientific knowledge'—a *falsity* is immediately added to this body if the finding is a mistaken discovery. By contrast, the effect of a missed discovery on this body is far less immediate.

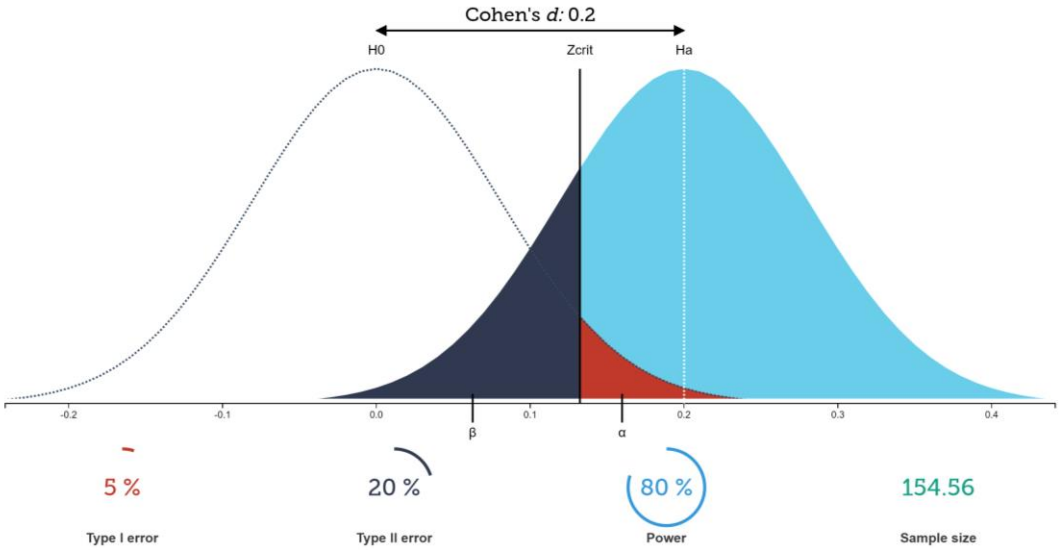

Fig. 1: The probability density distributions for H_0 (left) and H_1 (right) in a one-sided *t*-test for an effect size of $d = [(m_1 - m_0) / s] = .20$, given $\alpha = .05$, $(1 - \beta) = .80$, and N ≅ 155 per group (Source: https://rpsychologist.com/d3/nhst/, Kristoffer Magnusson, CC-BY license).

2.3 The body of scientific knowledge

A major assumption in NHST is that the body of scientific knowledge can neither be harmed, nor improved, if a researcher responds to a statistically *in*significant hypothesis test result ($P(H_0, D) > \alpha$) by maintaining H_0. This assumption provides a system-rational reason for a preference phenomenon known as *selective publishing*,[1]

[1] Selective publishing reflects Popper's (1959) *falsification principle* for theory choice: empirical hypotheses can be falsified but not verified. This principle makes statistically insignificant hypothesis test results uninformative in NHST because, while H_0 can in response to such results be maintained, these results cannot be interpreted as confirming H_0 (see Sect. 2.1). Statistically insignificant test results, therefore, often remain unpublished (*file drawer problem*; Rosenthal, 1979), making them harder to access when seeking to correct population effect size-estimates that a meta-analysis (which predominantly harvests published, statistically significant object-level test results) consequently overestimates (Rothstein, Sutton & Borenstein, 2005). Perhaps the best counter-measure against selective publishing is a combination of *results-blind manuscript evaluation*—where the decision to publish

aka a *publication bias* in favor of discoveries. Over time, this bias affects the shape of the body of scientific knowledge because most contributions to it will be statistically significant findings. Conversely, this body is unlikely to be shaped by statistically insignificant findings *regardless* of whether maintaining H_0 is a correct or incorrect response, i.e., whether nothing was to be discovered or something to be discovered was missed. A mistakenly maintained H_0 (β-error) can hence be compared to a safe bet. Whereas a mistakenly rejected H_0 (α-error) entails that what is "contributed" to this body is a false H_1, a falsity remaining in this body until future findings correct it. The epistemic risk of being misled by falsities thus makes a mistaken discovery (α-error) a *more serious* error than a missed discovery (β-error).

Therefore, Cohen may have reasoned, researchers are on the one hand epistemically justified to prefer an unpublished missed discovery over a published mistaken discovery. On the other hand, if the β-error rate exceeds .20, the entire field would in the long run "miss out" on too many missed discoveries. This on-balance reasoning offers an *epistemic* justification to avoid a body of scientific knowledge that includes too much of what is not there (α-error) and excludes too much of what is there (β-error).

While offering a prima facie plausible justification, the idea that resource restrictions limit the sample size a researcher can collect (Cohen, 1992, p. 156) appears to be a *practical* reason for $\beta = .20$. But this suggests that an epistemic reason is associated only with $\alpha = .05$. If so, then $\beta = .20$ as an acceptable long-run proportion of missed discoveries would be justified practically rather than epistemically. By Cohen's own standards, however, that sounds absurd. The primary sufficient justificatory reason for $\beta = .20$ as an upper error-bound can only be the epistemic reason to limit the long-run proportion of missed discoveries that fail to

---

a result occurring independently of its statistical significance increases the chance that insignificant results are published (Berlin & Ghersi, 2005; Chambers, 2013; Locascio, 2019)—and a *likelihood ratio hypothesis test*, which leaves statistically insignificant results informative to correcting population effect size-estimates and to (dis-)confirming H_0 (Krefeld-Schwalb, Witte & Zenker, 2018; Witte & Zenker, 2017).

inform the body of scientific knowledge. Resource restrictions, then, provide a *supererogatory* reason (i.e., an additional sufficient reason) for $\beta = .20$ as a lower error-bound. This latter reason, we argue, is *explanatory* rather than justificatory (Sect. 4).

An alternative reconstruction of the justificatory structure, we submit, would fail to account fully for the considerations Cohen brings to bear. This is distinct from claiming that Cohen's convention "gets it right."

3. The shape of the body of knowledge in behavioral science

3.1 Observed effect size and sample size

The statistical mark of the *replication* or *confidence crisis* is that published NHST-based studies in behavioral science normally report effect sizes that were observed under low statistical test power, corresponding to a larger β-error rate (Fletcher, 2021; Krefeld-Schwalb, Witte, Zenker, 2018; Szucs & Ioannidis, 2017a; Ioannidis, 2005; van Dongen & Sikorski, 2021; Wagenmakers et al., 2011). That studies in behavioral science yield publishable results despite being underpowered had already been recognized in Cohen (1962), who estimated the average statistical test power in the field, i.e., the $(1 - \beta)$-error rate, as a disappointing .18.

One reason for low statistical test power to arise is that published studies in behavioral science normally report *small* effects, defined as $d = .20$ (Cohen, 1965), that are observed in small samples (Cohen, 1962; 1992; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989; Szucs & Ioannidis, 2017b). The estimated median sample size of published studies in psychology, for instance, is $N = 40$ (Marszalek et al., 2011; Wetzels et al., 2011; see Bakker, van Dijk & Wicherts, 2012). Other things equal, however, only a large(r) sample allows observing a small effect under high(er) test power. To illustrate, we present the sample size for a one-sided *t*-test in Table 2. This explains why decreasing the β-error rate taxes a researcher's resources.

Table 2: The total minimum sample size in the experimental and control group for a one-sided *t*-test as a function of statistical test power $(1 - \beta)$ and effect size (d), given $\alpha = .05$.

	d			
$(1 - \beta)$	.01	.20	.50	.80
.40	38,726	97	15	6
.50	54,111	135	22	8
.80	123,651	309	49	19
.95	216,443	541	87	34

As suggested by subsequent similarly disappointing estimates of statistical test power (Cohen, 1992; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989), not only did behavioral science continue to rely on undersized samples given the observed effect sizes, but low median statistical test power also resulted from questionable ways of increasing the probability of obtaining a *publishable* finding, i.e., a statistically significant one. Besides the practice of *p*-hacking, for instance, "most studies involve tests of multiple hypotheses, [thereby] creating a gap between the power for any single test and the power for the collection of tests," wherefore despite every single test being underpowered, "the probability of rejecting at least one hypothesis in the collection of tests […] exceed[s] the probability that any specific hypothesis is rejected" (Maxwell, 2004, 148).

We return to the importance of statistical test power for discovery-oriented research below. What is clear already now is that researchers continued to treat the *p*-value, respectively the α-error rate, as (much) more important than the $(1 - \beta)$-error rate.

3.2 The p-value fallacy, statistical significance, and scientific importance

That samples are typically too small to yield well-powered test results holds for NHST-based research in the Fisher tradition, which recognizes only the *p*-value, as well as for research in the α- and β-error rate-recognizing Neyman-Pearson tradition, against the background of which Cohen advocated $\alpha = .05$ and $\beta = .20$.[2] A recent

[2] The *p*-value originates in the Fisher version of NHST. It states the probability of observing actual or more extreme data on the assumption that H_0 is true. The α-error

proposal to *abandon* NHST (Lakens et al., 2018; McShane et al., 2019; Trafimow et al., 2018)—because hypothesis tests that satisfy $p(D, H_0) < p = .05$ (or $p(D, H_0) < \alpha = .05$) are regularly interpreted as implying a probability > 95% that H_1 is true—lays much of the blame on the *p-value fallacy* (aka *prosecutor's fallacy*), an invalid inference reflecting the unwarranted transition from a probability to a likelihood, as well as "the mistaken idea that a single number [e.g., $p = .05$] can capture both the long run outcomes of a scientific study and the evidential meaning of a single result" (Goodman, 1999, p. 995; see Cohen, 1994, p. 997).

Yet, abandoning NHST is overly drastic. After all, avoiding the *p*-value fallacy requires no more than interpreting the *p*-value as the probability of observing an effect size equal to, or more extreme than, the observed effect size if H_0 is true. Indeed, nothing is problematic with the *p*-value itself, but rather with its well-documented misapplication and overinterpretation as a measure of evidence (Gómez-de-Mariscal et al., 2021; Halsey et al., 2015). This includes a quasi-mechanical identification of $p = .05$ with an observed effect size's statistical significance, the unwarranted transition from its statistical significance to its scientific importance, and a ritualistic practice of teaching a *p*-value, respectively an α-error rate, of 5% "[…] because it's what we do; [while] we do it because it's what we teach" (Wasserstein & Lazar, 2016, 129).

Already Fisher (1925), who proposed $p = .05$ as a conventional rejection criterion for H_0, had offered no more than a convenience justification for its specific value (Hubbard, 2016; Kennedy-Shaffer, 2019):

---

rate, originating in the Neyman-Pearson version of NHST, states the long-run chances of mistakenly rejecting H_0 (false positive) as the proportion of mistaken decision among all decisions to maintain/reject H_0. If the decision criterion is the *p*-value, then this probability is estimated based on data (objective interpretation), whereas if the decision criterion is the α-error rate, this probability is estimated based on a researcher's expected error rate (subjective interpretation). Because a sound evidence-based decision demands that the subjectively expected α-error rate is at least as large as the objective *p*-value, the conceptual differences between the *p*-value and the α-error rate are, in praxis, easily "hidden."

> The value [of the standard deviation] for which $p = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a [statistical] deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a negative [test-]result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, but no lowering of the standard of significance would meet this difficulty. (Fisher, 1925, p. 47; *notation adapted*)

In this way, $p = .05$ was from the outset accepted by convention. As Fisher observed, when the strength of association between two variables is determined by Pearson's (1900) χ^2-test, $p = .05$ roughly states the probability that an observed mean falls more than two standard deviations away from the mean of a normally distributed random variable. So, "we shall *not often be astray* if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real [rather than a mistaken] discrepancy" (Fisher, 1925, p. 79; *italics added*). As a threshold for rejecting H_0, then, already Fisher's (1925) statistical inference system (on which NHST is based) only offers a conventional justification for $p = .05$. Similarly, when Edgeworth coined the term 'statistical significance', in 1885, he merely wanted "a tool to indicate when a result warrants further scrutiny; [but] statistical significance was never meant to imply *scientific importance*" (Di Leo et al., 2020, p. 2; *italics added*; see Kennedy-Shaffer, 2019, p. 84).

3.3 Conventions by convention?

This makes it understandable why, despite the regular misinterpretation of its probability-based definition, a conventional acceptance of the *p*-value is convenient. But conventions hardly justify their application—a truism acknowledged not only by Cohen (1965) but also by Neyman and Pearson (1933), who are worth quoting in full:

> But whatever conclusion is reached, the following position must be recognized. If we reject H_0, we may reject it when it is true; if we accept

> H_0, we may be accepting it when it is false, that is to say, when really some alternative H_t [i.e., H_1] is true. These two sources of error can rarely be eliminated completely; in some cases, it will be more important to avoid the first, in others the second. We are reminded of the old problem considered by LAPLACE of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty? That will depend upon the *consequences of the error*; is the punishment death or fine; what is the danger to the community of released criminals; what are the current ethical views on punishment? From the point of view of mathematical theory all that we can do is to show how the risk of errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator. (Neyman & Pearson, 1933, p. 296; *italics added*)

While Neyman and Pearson task researchers themselves with striking the balance between both error rates, they too advocate $\alpha = .05$ so that "in the long run of experience, we shall not too often be wrong" (Neyman & Pearson, 1933, p. 291). Indeed, Neyman (1950, p. 262) himself suggested that α-errors are more serious than β-errors. To this, Cohen added the point-specific ratio $\alpha / \beta = 1 / 4$. But in doing so, the reasons for the "arbitrary but reasonable value" (Cohen, 1988, p. 56) $\beta = .20$ remained vague:

> First, I believe that generally the consequences of false positive claims (rejections of null hypotheses) are more serious than those of false negatives (acceptance of null hypotheses). This is in accord with the conventional scientific view of these matters. Present practice, which concerns itself solely with the former [i.e., the proportion of α-errors among published statistically significant test results], by ignoring the latter [i.e., the proportion of β-errors] implicitly treats them as if they were of no, or at least little, consequence. *My proposal maintains the usual emphasis but keeps the relation between the two risks within reasonable bounds*. Since the convention of the 5 per cent level for α has

> come to be generally used, my proposal implies a setting of a 'subjective general relative seriousness' of 20 per cent/5 per cent = 4. The second consideration, then, in setting the β risk convention of .20 is that it is consonant with a rough guess that type I errors are *in general* about four times as serious as type II errors. I would, of course, have no serious quarrel with anyone who claimed that the factor should be three or five (or even two o[r] six), but such is the nature of conventions. I offer this convention so diffidently because I would prefer to see [statistical test-] power values set ad hoc wherever possible. I deplore the slavish adherence to the quasi-official convention of 5 per cent for type I errors, which has resulted in its implicit equation with scientific truth for the positive claim and with respectability, if not ethical purity, for the claimant. But however abused, conventions have their use. (Cohen, 1965, [1]1958, 98f.; *italics added*)

While Cohen indicates—notice the reappearing term 'serious'—the lack of serious reasons to oppose conventions other than α / β = 1 / 4, he avoided offering a clear interpretation of 'general relative seriousness'. Instead, he put one convention on top of another. Because α = .05 was already conventionally accepted, he could advance the dependent convention β = .20. This is what recent scholarship simply reiterates. For instance, "in the internal dealings of science, errors of Type I [α] are in general regarded as *more problematic* than those of Type II [β]" because "those who claim the existence of an as yet unproven phenomenon have the burden of proof" (Hansson, 2018, p. 7; *italics added*). That is descriptively correct. But it, too, avoids offering reasons for allocating the burden of proof in this way.

3.4 Balancing error rates

The conventional α = .05 has been rightly criticized (e.g., Bakan, 1966; Benjamin et. al., 2017; Gigerenzer, 2018). One good reason to abandon a *fixed* α-error rate = .05 is that an observed mean satisfying $p < .05$ can be more likely as statistical test power increases if there is *no* effect than otherwise, even though $p > .05$ is *expected* to be less likely if there is an effect than otherwise (aka Lindley's paradox; see Maier & Lakens, 2022). That said, what proposals to further decrease *only* the α-

error rate (e.g., Bartos & Maier, 2022; Benjamin et al., 2018; Lakens et al., 2018) ignore is that "setting a blanket level of either 0.05 or 0.005, or anything else, forces researchers to pretend that the relative importance of Type I and Type II errors is constant" (Trafimow & Earp, 2017, 3; see Trafimow et al., 2018).

Perhaps "the real lesson we should take away from Cohen is to determine the relative seriousness of Type 1 and Type 2 errors and to balance both types of errors [*before* running a study, i.e.,] when a study is *designed*" (Maier & Lakens, 2022, p. 2; *italics added*). When faced with design choices, researchers understandably desire an *efficient* decision on H_0 and H_1. And, given limited resources dictate a fixed sample size, "it is typically possible to make decisions more efficiently by choosing error rates such that the combined cost of Type 1 and Type 2 errors is minimized" (Maier & Lakens, 2022, p. 3) by calculating the *weighted combined error rate* (WCE). Assuming prior probabilities for H_0 and H_1, as well as $P(H_0)$ and $P(H_1)$, WCE is calculated as:

$$\text{WCE rate} = \alpha\text{-error rate} \times P(H_0) + \beta\text{-error rate} \times P(H_1) \qquad (1)$$

For instance, if $P(H_0) = P(H_1) = .50$, plugging Cohen's convention into (1) yields $\text{WCE} = (.05 \times .50) + (.20 \times .50) = .125$, whereas symmetrical error rates yield $\text{WCE} = (.05 \times .50) + (.05 \times .50) = .050$, i.e., a combined error rate that is 2.5 times smaller.

Because all kinds of cost and benefit considerations can, given all kinds of prior probability assignments, influence how researchers balance the error rates (for discussion, see Lakens, 2022; Maier & Lakens, 2022), Cohen would assumedly have recognized that the relative seriousness of a β-error may exceed that of an α-error *in certain contexts*. For he writes, "[a]lthough the pure researcher cannot place a dollar utility value on the consequences of type II (and type I) errors, as can, for example, the industrial quality engineer, [they] can, by a subjective weighing of the consequences of an error in inference and the effort involved in producing data, approximate this approach" (Cohen, 1965, 98). Generally, the more serious the error type is, the less frequently one wants its token to occur. And yet, for discovery-oriented research, as resource restrictions limit the sample size a researcher can

collect and as mistaken discoveries are deemed more serious errors than missed discoveries, Cohen's default balance came to α / β = .05 / .20.

4. Inductive risk

4.1 Epistemic and non-epistemic considerations

By the mid-20th century, concurrent with developments in statistics and probability theory that demonstrated the ability to rigorously express the degree of confidence in a scientific hypothesis under test (Andersen & Hepburn, 2016, p. 25), the debate on hypothesis testing in the philosophy of science had suggested that understanding 'hypothesis testing' as a decision between possible actions requires acknowledging a *value* component (ibid.). Whether this component is what "drives" a decision to maintain or reject a hypothesis is today as controversial as whether such decisions require not only epistemic but also non-epistemic, practical considerations.

Epistemic considerations (e.g., simplicity, explanatory power, or predictive accuracy) are associated with the truth-likeness of a hypothesis (e.g., Kuhn, 1962; 1977), whereas non-epistemic or practical considerations (e.g., moral, legal, or social goods relevant to public policy) are associated with the *utility* of a scientific result. If both kinds of considerations inform the general relative seriousness of errors, and thus the balance between α- and β-errors, then although

> […] science gives higher priority to avoiding type I errors [α-errors] than to avoiding type II errors [β-errors], the balance can shift when errors have practical consequences. This can be seen from a case in which it is uncertain whether there is a serious defect in an airplane engine. A type II error, i.e., acting as if there were no such a defect when there is one, would in this case be counted as more serious than a type I error, i.e., acting as if there were such a defect when there is none. (Hansson, 2018, p. 7)

Examples that shift the balance towards preferentially avoiding a β-error are easy to find (e.g., concerning legal cases, the environment, or public health). But these examples are *orthogonal* to Cohen's convention, the seeming reasonableness

of which, we argued, is owed primarily to epistemic rather than practical considerations.

Of course, a practical consideration such as the utility of a hypothesis is what Cohen may recur to *implicitly*, a speculation gaining initial plausibility from debates between, among others, Fisher and Neyman and Pearson (Howie, 2002; Lenhard, 2006; Marks, 2000). For Fisher (1955), who understood 'testing a hypothesis' as applying a method to decide whether H_0 can be accepted as true, the truth of H_0 counts more than its utility. In his view, even a true H_0 should be rejected when evidence consistent with it is scant compared to evidence consistent with an equally plausible alternative hypothesis. Whereas Fisher viewed significance tests and *p*-values as continuous measures of evidence against the truth of H_0, Neyman and Pearson addressed the question of whether a researcher should *act* as if H_x is true (Neyman, 1956; Pearson, 1955). While they acknowledge, as we saw, that an evidence-based decision to maintain/reject H_x must be sensitive to both kinds of error, for them, unlike for Fisher, this decision also depends on a hypothesis' utility.

Cohen thus agrees with Neyman and Pearson that α-errors are more serious than β-errors. Moreover, he agrees with Fisher that a hypothesis test (rather than aiming to maximize a decision's utility) aims at determining a hypothesis's truth and that even a true hypothesis ought to be rejected if evidence consistent with it is scant. This latter agreement may suggest that utility considerations are implicit in Cohen's notion of *evidence*. But even if so, Cohen's primary reason for α / β = .05 / .20 remains an epistemic consideration. In justifying his convention, then, utility considerations do not appear to play a load-bearing role.

4.2 The functional role of risk-related information

The functional role of the probabilities associated with each error type implies that hypothesis-related decisions are risk-related (Hansson, 2018). The first kind of risk concerns the *veracity* of information included in the body of scientific knowledge. Limiting this risk is the purpose of proof standards.

In the current standard model of physics, for instance, the five-sigma ($5 \times \sigma$) proof standard—implying that H_0 is rejected only if an observed mean deviates by at least five standard deviations from a *theoretically expected* mean (H_1)—corresponds to an α-error rate of .000003. So, if H_0 is true, a similarly small deviation is expected

only roughly once in three million tests (Bird, 2018, p. 17). While, given this standard, a mistaken discovery is very unlikely to enter the body of scientific knowledge in physics, the broad absence of predictive theories in behavioral science implies that observed means are here tested against chance (vs. a theoretically expected mean). And the conventional α-error rate of .05 (or a proof standard of $1.96 \times \sigma$; Bentley, 2021, p. 2) implies that a mistaken discovery occurs about once in 20 tests.

The second kind of risk concerns the *utility* of hypothesis-related decisions. A paradigm example is to diagnose a healthy person as diseased, or *vice versa*. Compared to the risk of contributing a mistaken discovery to the body of scientific knowledge, the seriousness of a mistaken medical diagnosis can—if greater negative utility is assigned to a β-error than to an α-error—imply a change in the functional role of risk-related information. For instance, assume a reliable test to diagnose person *X* as free from a potentially fatal contagious infection *I*.[3] Cohen's convention would state that the seriousness of *X* not having *I*, given the test says *X* has *I* (α-error), exceeds that of *X* having *I*, given the test says *X* does not have *I* (β-error). But this cannot be right. Someone mistakenly diagnosed as non-infected presents a risk of spreading *I*, a risk someone mistakenly diagnosed as infected cannot present. In the false positive case, *S* may self-quarantine and become bored—no doubt a mild

[3] For example, the reliability of administering the *Reverse-Transcription Polymerase Chain Reaction* (RT-PCR) test, a common diagnostic test for SARS-COVID-19, varies with laboratory conditions and the kind of polymerase used. The test's error rates under *real* conditions (vs. test validation conditions) are variously estimated as $.03 < \alpha < 3.0$ and $.09 < \beta < .19$ (Arevalo-Rodrigez et al., 2020; Cohen, Kessel & Milgroom, 2020; Long et al., 2020). Even if the true positive rate is at its peak level—such that test-sensitivity (the long-run rate of true positive over true positive plus false negative test results) is maximal—one should expect $.167 < \alpha < .29$, while $\beta = .21$ (ibid.). The large β-error rate implies that, while the joint β-error rate of two (or more) independent RT-PCR tests decreases with the number of tests, a *single* RT-PCR test is insufficient "to 'clear' people as being non-infected" (Bentley, 2021, 9).

negative consequence. But in the false negative case, the consequences may be tragic.

The relative seriousness of both errors thus favors preferentially avoiding a missed discovery (β-error) in the context of public health. While the above example of an *individual* medical test result takes Cohen's convention out of its intended context, a reasoned preference to preferentially avoid β-errors rather than α-errors also obtains when medical tests are administered at the *population* level, or when treatments are developed for a population that needs them—cases that fall squarely within the intended context of Cohen's convention. Similar contexts point back at the question, considered by Laplace, of how many votes in a court of judges are needed for a conviction. As Neyman and Pearson acknowledge, the disutility of a mistaken verdict, test result, or diagnosis may vary (e.g., between boredom and death). In balancing error rates, then, ignoring such practical considerations would be poor advice.

4.3 The argument from inductive risk

If practical considerations cannot be ignored, and particularly if the error type to be preferentially minimized has policy implications, then a hypothesis-related decision should acknowledge the *value-laden* character of science (Diekmann & Peterson, 2013; Lemons et al., 1997). This idea runs counter to the ideal of value-free science, stating that the justification of scientific knowledge should be free of non-epistemic considerations (Betz, 2016), an ideal that various authors (e.g., Douglas, 2009; John, 2016; Rudner, 1953) object to using the *argument from inductive risk*, seeking to show that a balancing of α- and β-errors requires appealing to both epistemic and practical considerations.

The Argument from Inductive Risk (John, 2016, p. 3)

1. Scientists accept or reject hypotheses.
2. Hypotheses typically fail to be deductively entailed by the available evidence.
3. Scientists face 'problems of inductive risk': they risk accepting false hypotheses (false positive errors) or rejecting true hypotheses (false negative errors).
4. A determination of the trade-off between the two error types must appeal to non-epistemic considerations associated with the consequences of these errors.

Therefore, scientific inference must appeal to non-epistemic considerations.

Scientific inference primarily demands an *epistemic* standard. The more stringent this epistemic standard is—i.e., the more evidence of a specific kind is needed to decide on a hypothesis—the less likely scientists are to reject truths or maintain falsehoods. Of course, a stringent epistemic standard is no less appropriate if a hypothesis-related decision is sensitive to the disutility of error. Therefore, the claim that non-epistemic considerations are *indispensable* in scientific inference is plausible, only if striking a balance between error rates requires an appeal to non-epistemic considerations—exactly as the argument from inductive risk suggests.

While the argument's first three premises are widely accepted, the fourth raises suspicion. Would not the appeal to non-epistemic considerations lead away from the truth, predictive accuracy, or even logical consistency, thus creating a *bias* (Hudson, 2022, 211)? While biases (e.g., anchoring, overconfidence, or confirmation bias) need not entail that scientific inference is value-laden, nor *vice versa* (Douglas & Elliot, 2022, p. 202), only an appeal to non-epistemic considerations can *explain* why, in some contexts of inductive risk (e.g., in medical diagnosis), a β-error is more serious than an α-error. Non-epistemic considerations can thus be indispensable to explain a specific error-rate balance.

While the greater seriousness of a β-error provides a sufficient reason to reject Cohen's convention outside of the context of discovery-oriented research, a sufficient reason to reject it within this context is the collectively unreasonable influence it has exerted on behavioral science.

5. Long-run epistemic consequences

5.1 The replication probability of a true observed effect size

As systematic attempts to replicate a representative sample of published observed effect sizes have broadly failed (e.g., Many Labs Projects 1-5; Ebersorle et al., 2020), most effect size estimates published in behavioral science journals are presumably best thought of as *probably non-replicable*. This status, we claim, is (at least partially) a consequence of the collective unreasonable influence of Cohen's convention on behavioral science. This influence counts because probably non-

replicable effect size estimates are unsuitable for theory construction research, thus dampening the prospects for developing a progressive science of human behavior.

The replication probability of a *true* observed effect size is determined largely by the statistical test power (i.e., the $(1 - \beta)$-error rate) of the study observing it. (The hedge 'largely' derives from a regression effect, inversely correlated with sample size, making it more likely than not that a replication study observes an effect *slightly smaller* than in the initial study (Fiedler & Prager, 2018).) A study's $(1 - \beta)$-error rate is a function of its α-error rate, the observed effect size, d, and the sample size, N. Even where N is determined primarily by a researcher's resources, the observed effect size (d) is best thought of as being determined primarily by "how the world is." Other things equal, if N increases, so does the $(1 - \beta)$-error rate, i.e., the β-error rate decreases. Specifically, given constant d, increasing N decreases the α- and β-error rates *symmetrically*, while, given constant d and only a single fixed error rate (e.g., α), increasing N decreases only the other error rate (e.g., β).

A convention of the form $\alpha < \beta$ thus implies that, if d and N are constant, then the β-error rate never matches the p-value, respectively the α-error rate. This has consequences for the veracity of the effect size estimates that researchers contribute to the body of scientific knowledge. If a statistically significant *true* effect size $d = x$ is initially observed under $(1 - \beta) = .80$, as Cohen's convention suggests, then the long-run chance of re-observing x in independent replication studies is 80%. So, 80 out of 100 replication studies would succeed, and 20 would fail. That 80 : 20 proportion is what Cohen must have found acceptable, presumably *because* he had estimated the average observed statistical test power to be a disappointing $(1 - \beta) = .18$ (Cohen, 1962). For all we can assume, the laudable intention behind his convention was for behavioral science studies to achieve *at least* $(1 - \beta) = .80$.

5.2 Statistical test power, heterogeneity, and theory construction

However, in the largest behavioral science field, psychology, the median observed statistical test power of published effect sizes is estimated as $(1 - \beta) = .35$ (Bakker et al., 2012), massively undermining Cohen's convention (Christopher, 2019; Open Science Collaboration, 2015; Stanley et al., 2018). More precisely, the estimated average statistical test power of published psychological studies that report small ($d = .20$), medium ($d = .50$), and large ($d = .80$) effect sizes is estimated as,

respectively, $(1 - \beta) = .23, .62, .84$ (Thorn et al., 2019, 13). An earlier estimate for observed effect sizes published in cognitive neuroscience and psychology comes to $(1 - \beta) = .12, .44, .73$ (Szucs & Ioannidis, 2017b). What approximates the 80 : 20 proportion of probably replicable effects suggested by Cohen's convention thus are *large* observed effects.

But even large observed effects that feature $(1 - \beta) \approx .80$ cannot automatically improve the prospects for a progressive science of human behavior. In behavioral science, after all, meta-analytically estimated effect sizes describe a specific pattern: they are either *small* and *homogenous* or *large(r)* and *highly heterogeneous* (Linden & Hönekopp, 2021; 366, Fig. 5; see Olsson-Collentine, Wicherts & van Assen, 2020; Schauer & Hedges, 2020). The heterogeneity of a meta-analytical effect size estimate measures the degree to which (topically related) object-level observed effect size estimates vary around the meta-analytical estimate. A *high* degree of heterogeneity indicates that the underlying object-level effects differ vastly in size, and a *low* degree of heterogeneity indicates that the underlying object-level effects are similar in size.

This pattern suggests that a *clearly* observed object-level effect size in behavioral science translates into a small meta-level effect size, whereas a large(r) meta-level effect size is underlain by *diffuse* object-level observations. Given what meta-analytical research makes available, then, this pattern keeps from identifying population effect sizes that are worthwhile parameters for theory construction research. After all, that a *homogenous* population effect size is normally *small* means that such an effect explains little more of the observed variance (r^2) than is "explained" by a random effect ($d = 0$). Given effect sizes $d < .45$, for instance, we find $r^2 < .05$ % (Cohen, 1977, Table 2.2.1). So, particularly a *very small* population effect size can be accounted for by the measurement error alone, leaving it unclear whether there *is* an effect.

This makes small effects poor parameters for theory construction. Conversely, if a *large(r)* population effect is normally highly *heterogeneous*, then the process of theoretically modeling it remains subject to vast uncertainty because it remains unclear *exactly which* effect size it is that a valid theoretical construct would model. Such effects, then, are equally poor parameters for theory construction. Adding gravity to this problem is that recent critiques of meta-analytical and

replication methods strongly suggest that, because of publication bias, population effect size estimates in behavioral science do anyway state *overestimates* (Klein et al., 2018; Schäfer & Schwarz, 2018).

5.3 A progressive science of human behavior

While a progressive science of human behavior thus depends on recognizing the *precise identification of replicable effect size estimates of sufficient size* as the ultimate goal of discovery-oriented research, the most immediate way of obtaining more precise object-level observations is to collect larger samples (*law of large numbers*). Of course, even if very similar effect sizes are observed in a series of only $i = 3$ independent replication studies under $(1 - \beta) = .80$, the joint statistical test power of the entire series, $(1 - \beta)^i$, registers already close to chance ($.80^3 = .51$) (Francis, 2012). Obtaining a well-powered series ($(1 - \beta) > .95$) thus requires that each study in this series achieves maximal statistical test power—again requiring a large *N*. While the $(1 - \beta)^i$ measure strictly counts for theory *confirmation*, what counts for theory *construction* is that observed effect sizes across the entire series are *similar*. Alas, evaluating their similarity equally requires a precise effect size estimate, and that estimate's precision increases with *N* (Witte, Stanciu, Zenker, 2022). So, come what may, large(r) samples are needed.

This need, however, is precisely what the $\beta = .20$ part of Cohen's convention fails to reflect, and what behavioral science studies normally fail at when statistical test power is consistently too low. Indeed, as individual scientists typically publish an initially observed effect size $d = x$ under a "tight" α-error rate but a "lax" β-error rate, they leave assessing its replicability to their colleagues, "outsourcing" what a progressive behavioral science *as a field* cares about the most. This may appear to be good scientific practice, too, because for *x* to become a serious candidate for theory construction, colleagues must anyway report independent re-observations not just of the effect's direction, but its size.

Yet this practice is doomed to fail if, as is typical in behavioral science, *x* is initially observed (and published) under low statistical test power. We saw that already $(1 - \beta) = .80$ entails only an 80% probability of replication success. The fact that a typical study's $(1 - \beta)$-error rate is *much smaller* thus translates into a typically low probability of replication success. Therefore, a colleague would

typically *lack* rational reasons *to even attempt* replicating *x*, because a replication failure is too probable. That the probability of a replication success in behavioral science is typically low thus helps to understand why the field experiences a replication crisis in the first place. (Based on the median observed statistical test power of published effects, the crisis could have been (fallibly) predicted.)

Without committing to *small* and *even* error rates ($\alpha = \beta << .05$), and thus to using large(r) samples, then, it is hard to see how researchers might obtain the independently reproducible effect size estimates of sufficient size that are *discoveries*, discoveries on which the development of a progressive science of human behavior depends.

6. Objections

6.1 Overview

We claimed that the replication crisis in behavioral science results in large part from publishing statistically significant effect sizes that are observed under low statistical test power, that this practice dampens the prospects for the development of a progressive science of human behavior, and that the practice itself reflects the asymmetrical error rates of Cohen's convention. Key challenges to this claim pivot on the role of irreversible experimental units in different types of replication studies, the efficiency of scientific inquiry, the base rate of replicable hypotheses, and the remedial potential of meta-analysis.

6.2 Replication types and irreversible experimental units

Whereas *exact* replications identically operationalize an initial study (e.g., in the same lab, the next day), *direct* replications only operationalize aspects thought to be causally relevant to a finding (e.g., in another lab, one month later), and *conceptual* replications operationalize entirely different aspects while addressing the same theoretical concept or hypothesis (Hudson, 2023; Matarese, 2022). A conceptual replication can thus draw on different samples or operationalizations to manipulate the behavioral responses of what are in principle *irreversible* experimental units (e.g., people, social groups, systems). If, despite such differences, the observed effect size is sufficiently similar to that observed in an initial study, the effect is more likely to be a true positive than not (Crandall & Sherman, 2016).

In the case of exact and direct replications, the α- and β-error rates are intended to apply to a series of replication studies only if (i) each study samples randomly from the same population and (ii) identically operationalizes all aspects of an initial study, respectively all causally relevant aspects (Neyman, 1937, pp. 334-335; Neyman & Pearson, 1928, esp. pp. 177, 231, 232; see Rubin, 2019). While most behavioral science studies rely on a convenience sample—making randomized sampling procedures *counterfactually* assumed—the second condition means that the set of causally relevant aspects "reflects current beliefs about what is needed to [re]produce a finding" (Nosek & Errington, 2017, p. 1). Thus, a false belief about what is needed, the influence of moderators, or a sample-specific lack of sensitivity, may explain why a direct replication fails. And, in virtue of dealing with irreversible experimental units, *direct* replication is all that behavioral science can achieve in the first place (Rubin, 2019).

But to therefore resort, as Rubin (2019) argues, to the Fisherian *sample-specific* concept of error rather than to Neyman-Pearson's concept—interpreted "in relation to a series of samples that could have been randomly drawn from the exact same null population" (Rubin, 2019, p. 5816)—would render *all* observed effects in behavioral science *contingently* replicable. As experimental units, after all, "people are time- and context-sensitive units of analysis that have the potential to interpret identical situations in multiple different ways (Rubin, 2019, p. 5812). A finding that replicates would thus be as unnewsworthy as the opposite outcome.

Like in the case of the explanatory relevance of potential moderators, however, the relevance of irreversible experimental units cannot be established by argument but must be modeled theoretically and demonstrated experimentally. Until then, Rubin's (sophisticated) explanation offers an *apology* for failed replications in behavioral science.

6.3 The efficiency of scientific inquiry

Unlike researchers who take the replication crisis not to pose a serious problem for behavioral science (e.g., Redish et al., 2018), Lewandowsky and Oberauer (2020) argue that low replicability may reduce the cost of producing scientific knowledge while increasing its efficiency if questionable research practices (QRPs) were abandoned in an *idealized* transparent scientific community. In this community,

either "individual studies are published and are replicated after publication, but only if they attract the community's interest" (ibid.)—itself equated with a citation pattern of a published study modeled by fitting an actual citation pattern in psychology (ibid. p. 10)—or "all findings are replicated before publication to guard against replication failures" (ibid., p. 1).

Using simulations, Lewandowsky and Oberauer (2020) show that, compared to the first replication regime, the second "incurred an additional cost of around ten studies […] [,] represent[ing] ~10% of the total effort the scientific community expended on data collection" (ibid., p. 4). Although the "analysis of replicability confirms that citations do not predict replicability" (ibid.), the authors suggest that, regardless of the replication regime, "the probability of replication of a study *increases* with the number of citations" (ibid., p. 3; *italics added*). In discovery-oriented research, then, the efficiency of generating scientific knowledge would be proportional to the citation counts of published studies.

While this may sound encouraging, the simulated generation of scientific knowledge under ideal conditions (without QRPs) is trumped by actual conditions. And, in top psychology and economics journals, cited more frequently are published studies that report non-replicable effects (Sena-Garcia & Gneezy, 2021)—the exact opposite pattern of what the simulations suggest. Formally, moreover, well-powered observed effect sizes ($\alpha = \beta < .05$) are less likely to be β-errors than similarly-sized but underpowered observed effects ($\alpha < .05,\ \beta > .05$) (Witte, Stanciu & Zenker, 2022). So, even without QRPs, statistical test power remains crucial for a progressive science of human behavior.

6.4 The base rate of false hypothesis

Indirectly arriving at the same conclusion, Bird (2018) acknowledges publication bias and QRPs as exasperating the replication crisis, yet observes that a large proportion of failed direct replications can be consistent with high-quality science. If "the field of science in question produces a high proportion of false hypotheses prior to testing" (ibid., p. 1)—Bird stipulates a base rate of 90% of false hypotheses—then $\alpha = .05$ and a well-powered hypothesis test—Bird stipulates $(1 - \beta) = .95$—would nevertheless let 1/3 of statistically significant (published) findings be false positives. As this 1/3 "survives" testing, it would show up as failed re-tests, i.e., as failed direct

replications. While formally correct, "driving" this explanation of replication failures in behavioral science is a *stipulated* base rate of false hypotheses. The true base rate, however, is highly uncertain. Pointing once again to the *file-drawer problem* (Rosenthal, 1979), estimating the true base rate depends on estimating the proportion of *unpublished* studies that correctly maintain H_0. But this proportion is highly uncertain, too.

In his conclusions, Bird agrees with Fisher (1934, [1]1925, p. 123) that "confidence to be placed in a result depends not only on the magnitude of the mean value obtained, but equally on the agreement between parallel experiments." And for this agreement to be assessed properly—as Bird would acknowledge irrespective of the base rate of false hypotheses—requires *precise* effect size estimates, featuring the very small and even error rates ($\alpha = \beta << .05$) that only large samples can offer.

6.5 Meta-analysis to the rescue?

Given that object-level effect size estimates are underpowered, it may appear compelling that meta-analytical procedures (Hunter & Schmidt, 2004; Stanley et al., 2018) could aggregate a large number of (topically related) studies to produce better effect size estimates (Fletcher, 2022). However, the quality of a meta-analytical effect size estimate primarily depends on the quality of the underlying object-level estimates. Given publication bias, the quality criterion is the *precision* of a published object-level estimate, which increases with *N*. And, precision is precisely what is lacking when published object-level effect size estimates are underpowered. Indeed, modeling results suggest that 90% of published object-level effect size estimates be *discarded*, to meta-analytically estimate an effect size by "averag[ing] the most precise 10% of the reported [object-level] estimates" (Stanley et al., 2010, p. 1). The importance of statistical test power, therefore, also holds in meta-analysis.

7. Conclusion

In 1965, as Jacob Cohen advanced the convention that behavioral scientists conducting discovery-oriented research adopt default error rates that mirror the *general relative seriousness* of an error type, he coordinated the antecedently accepted $\alpha = .05$ to $\beta = .20$, i.e., $\alpha / \beta = 1 / 4$. Doing so, we argued, he sought to *decrease* a far larger β-error rate that was then characteristic of behavioral science

studies. In effect, however, his convention made it acceptable that behavioral science studies came to rely on *strongly uneven* error rates of the form $\alpha << \beta$.

Cohen's primary sufficient reason for $\alpha / \beta = 1 / 4$, we argued, was epistemic: to limit the proportion of mistaken discoveries that a published false finding would add to the body of scientific knowledge while limiting the proportion of missed discoveries that would fail to inform this body. By contrast, his supererogatory reason for $\beta = .20$—that resource restrictions normally keep researchers from collecting larger samples (that yield lower β-error rates)—was practical. Because this practical reason is indispensable to *explain* the balance of errors that Cohen proposed, it is easily mistaken for the justificatory reason that, we argued, it is not.

Cohen's convention thus offers an epistemic reason for a missed discovery (itself unlikely to be published) to be preferred over a mistaken discovery by an individual researcher. For an entire field, however, the convention is collectively unreasonable because, in the long run, $\alpha << \beta$ entails insufficient statistical test power, i.e., a low probability that published effect size estimates are replicable, thus dampening the prospects for a progressive science of human behavior. Without committing to *small* and *even* error rates ($\alpha = \beta << .05$), and thus to large(r) samples, then, it is hard to see how one might obtain the independently reproducible effect size estimates of sufficient size that are discoveries, discoveries on which the development of a progressive science of human behavior depends.

Acknowledgments

The authors have no conflict of interest to declare. A.A. wrote a draft that both authors jointly developed. F.Z. edited the final submitted version. We thank two reviewers for this journal, one of whom identified as F.J.W. Oude Maatman, for their comments that helped us to improve an earlier version of this manuscript. We also express our gratitude to audience members at the following conferences and events: EPSA 2023 (Belgrade, Serbia), LICPOS 2023 (Lisbon, Portugal), ENPOSS22 (Málaga, Spain), the workshop "Scientific progress – individual and collective" (Amsterdam, The Netherlands), and a research group meeting at GESIS (Mannheim, Germany). Finally, we extend our thanks to the members of the Philosophy Department and the Cognitive Science Group at Boğaziçi University,

Istanbul, Türkiye; the MTR research group (funded 2019-2022 by TÜBİTAK under grant 118C257, the support of which both authors gratefully acknowledge); and the following individuals: Samuel Fletcher, Gregor Garbor, Esra Mungan, Slobodan Perovic, Bican Polat, Ljiljana Radenovic, Patricia Rich, Bartłomiej Skowron, Adrian Stanciu, Jan Albert van Laar, Erich H. Witte, and Zsófia Zvolenszky.

References

Andersen, H.; & Hepburn, B. (2015). Scientific method. In Zalta, E.N. (ed.). *The Stanford Encyclopedia of Philosophy* (Summer 2016). https://plato.stanford.edu/archives/sum2016/entries/scientific-method/

Cohen, A.N.; Kessel, B.; & Milgroom, M.G. (2020). Diagnosing SARS-CoV-2 infection: the danger of over-reliance on positive test results. *medRxiv Preprints*, PPR157234. https://doi.org/10.1101/2020.04.26.20080911

Arevalo-Rodriguez, I.; Buitrago-Garcia, D.; Simancas-Racines, D.; et al. (2020). False-negative results of initial RT-PCR assays for COVID-19: A systematic review. *PloS one*, 15(12), e0242958. https://doi.org/10.1371/journal.pone.0242958

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437.

Bakker, M.; van Dijk, A.; & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554. https://doi.org/10.1177/1745691612459060

Bartoš, F.; & Maier, M. (2022). Power or alpha? The better way of decreasing the false discovery rate. Metapsychology, 6, MP.2020.2460. https://doi.org/10.15626/MP.2020.2460

Bentley P.M. (2021). Error rates in SARS-CoV-2 testing examined with Bayes' theorem. *Heliyon*, 7(4), e06905. https://doi.org/10.1016/j.heliyon.2021.e06905

Benjamin, D.J.; Berger, J.O.; Johannesson, M.; et al. (2018). Redefine statistical significance. *Nature Human Behavior*, 2, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berlin, J.A.; & Ghersi, D. (2005). Preventing publication bias: registries and prospective meta-analysis. In: Rothstein, H., Sutton, A, & Borenstein, M.

(eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments* (pp. 145–174). London: John Wiley & Sons, Ltd.

Betz, G. (2013). In defence of the value free ideal. *European Journal for Philosophy of Science*, 3, 207–220. https://doi.org/10.1007/s13194-012-0062-x

Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, 72(4), 965–993. https://doi.org/10.1093/bjps/axy051

Chambers, C.D. (2013). Registered Reports: A new publishing initiative at Cortex [Editorial]. *Cortex*, 49(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Christopher R.B. (2019). Effect size guidelines, sample size calculations, and statistical power in Gerontology, *Innovation in Aging*, 3(4), *igz036*. https://doi.org/10.1093/geroni/igz036

Crandall, C.S.; & Sherman, J.W. (2015). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, 66, 93–99. http://dx.doi.org/10.1016/j.jesp.2015.10.002

Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145–153.

Cohen, J. (1965). Some statistical issues in psychological research. In: B.B. Wolman (ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1970). Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement*, 30(4), 811–831.

Cohen, J. (1977, [1]1969). *Statistical Power Analysis for the Behavioral Sciences* (revised edition). London: Academic Press.

Cohen, J. (1988; [1]1969). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah: L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. https://doi.org/10.1037/0003-066x.49.12.997

Cohen, A. N.; Kessel, B.; & Milgroom, M. G. (2020). Diagnosing SARS-CoV-2 infection: the danger of over-reliance on positive test results *MedRxiv Preprint*, September 28, 2020. https://doi.org/10.1101/2020.04.26.20080911

Di Leo, G.; & Sardanelli, F. (2020). Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European Radiology Experimental*, 4(1), 1–8.

Diekmann, S.; & Peterson, M. (2013). The role of non-epistemic values in engineering models. *Science and Engineering Ethics*, 19, 207–218. https://doi.org/10.1007/s11948-011-9300-4

Díez, J.A. (2011). On Popper's strong inductivism. *Studies in History and Philosophy of Science Part A*, 42 (1), 105–116.

Douglas, H.E. (2009). *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Douglas, H.; & Elliott, K. C. (2022). Addressing the reproducibility crisis: A response to Hudson. *Journal for General Philosophy of Science*, 53, 201–209. https://doi.org/10.1007/s10838-022-09606-5

Ebersole, C.R.; Mathur, M.B.; Baranski, E.; et al. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, 3(3), 309–331. https://doi.org/10.1177/2515245920958

Edwards, W.; Lindman, H.; & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242. http://dx.doi.org/10.1037/h0038674

Edwards, A.W.F. (1972). *Likelihood.* Cambridge: Cambridge University Press.

Fiedler, K.; & Prager, J. (2018). The regression trap and other pitfalls of replication science—illustrated by the report of the Open Science Collaboration. *Basic and Applied Social Psychology*, 40(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh and London: Oliver and Boyd.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. New York: Hafner.

Fienberg, S.E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1), 1–40.

Fletcher, S.C. (2021). The role of replication in psychological science. *European Journal for Philosophy of Science*, 11, 23. https://doi.org/10.1007/s13194-020-00329-2

Fletcher, S.C. (2022). Replication is for meta-analysis. *Philosophy of Science*, 89(5), 960–969. https://doi.org/10.1017/psa.2022.38

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, 7, 585–594.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198-218. https://doi:10.1177/2515245918771329

Gómez-de-Mariscal, E.; Guerrero, V.; Sneider, A.; et al. (2021). Use of the p-values as a size-dependent function to address practical differences when analyzing large datasets. *Scientific Reports*, 11, 20942. https://doi.org/10.1038/s41598-021-00199-5

Goodman, S.N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12), 995. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Halsey, L.; Curran-Everett, D.; Vowler, S.; & Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12, 179–185. https://doi.org/10.1038/nmeth.3288

Hansson, S.O. (2018). Risk. In: Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 Edition). https://plato.stanford.edu/archives/fall2018/entries/risk/

Henkel, R.E.; & Morrison, D.E. (eds) (1970). *The Significance test controversy: a reader.* Aldine: Transaction Publishers.

Howie, D. (2002). *Interpreting probability: controversies and developments in the early twentieth century*. Cambridge: Cambridge University Press.

Hubbard, R. (2015). *Corrupt research: The case for reconceptualizing empirical management and social science*. London: Sage Publications.

Hudson, R. (2022). Rebuttal to Douglas and Elliott. *Journal for General Philosophy of Science*, 53, 211–216. https://doi.org/10.1007/s10838-022-09616-3

Hudson, R. (2023). Explicating exact versus conceptual replication. *Erkenntnis*, 88, 2493–2514. https://doi.org/10.1007/s10670-021-00464-z

Hunter, J. E.; & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Los Angeles: Sage Publications.

Ioannidis, J.P.A. (2005). Why most published research findings are false. *PloS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

John, S. (2016). From social values to p-values: The social epistemology of the intergovernmental panel on climate change. *Journal of Applied Philosophy*, 34(2), 157–171. https://doi.org/10.1111/japp.12178

Kennedy-Shaffer, L. (2019). Before p< 0.05 to beyond p< 0.05: using history to contextualize p-values and significance testing. *The American Statistician*, 73(sup1), 82–90. https://doi.org/10.1080/00031305.2018.1537891

Klein, R.A.; Vianello, M.; Hasselman, F., et al. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. https://doi:10.1177/2515245918810225

Krefeld-Schwalb; A., Witte, E.H.; & Zenker, F. (2018). Hypothesis-testing demands trustworthy data—a simulation approach to inferential statistics advocating the research program strategy. *Frontiers in Psychology*, 9, 460. https://doi.org/10.3389/fpsyg.2018.00460

Kovacs, M.; van Ravenzwaaij, D.; Hoekstra, R.; & Aczel, B. (2020). SampleSizePlanner: A tool to estimate and justify sample size for two-group studies. *Advances in Methods and Practices in Psychological Science*, 5(1). https://doi.org/10.1177/25152459211054059

Krüger, L.; Gigerenzer, G.; & Morgan, M.S. (1987). *The probabilistic revolution.* Boston: The MIT Press.

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago: Chicago University Press.

Kuhn, T.S. (1977). Objectivity, value judgment and theory choice. In: Kuhn, T.S., *The essential tension: Selected studies in the scientific tradition and change* (pp. 356–367). Chicago: University of Chicago Press.

Lakens, D.; Adolfi, F.G.; Albers, C.J.; et al. (2018). Justify your alpha. *Nature Human Behavior*, 2, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1): 33267. https://doi.org/10.1525/collabra.33267

Lemons, J.; Shrader-Frechette, K.; & Cranor, C. (1997). The precautionary principle: Scientific uncertainty and type I and type II errors. *Foundations of Science*, 2, 207–236. https://doi.org/10.1023/A:1009611419680

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science*, 57(1), 69–91. https://doi.org/10.1093/bjps/axi152

Lewandowsky, S.; & Oberauer, K. (2020). Low replicability can support robust and efficient science. *Nature Communications*, *11*(358), 1-11. https://doi.org/10.1038/s41467-019-14203-0

Linden, A.H.; & Hönekopp, J. (2021). Heterogeneity of research results: A new Perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, 16(2), 358–376. https://doi.org/10.1177/1745691620964193

Locascio, J.J. (2019). The impact of results blind science publishing on statistical consultation and collaboration. *The American Statistician*, 73(suppl 1), 346–351.https://doi.org/10.1080/00031305.2018.1505658

Long, Q. X.; Liu, B. Z.; Deng, H. J.; et al. (2020). Antibody responses to SARS-CoV-2 in patients with COVID-19. *Nature Medicine*, 26(6), 845–848. https://doi.org/10.1038/s41591-020-0897-1

Longino, H. (2020). Afterward: Data in Transit. In: S. Leonelli and N. Tempini (eds.), *Data Journeys in the Sciences* (pp. 391–400). Cham: Springer.

Maier, M.; & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, 5(2). https://doi:10.1177/25152459221080396

Marks, H.M. (2008). *The progress of experiment: science and therapeutic reform in the United States, 1900-1990*. Cambridge: Cambridge University Press.

Marszalek J.M.; Barber C.; Kohlhart, J.; & Holmes, C.B. (2011). Sample size in psychological research over the past 30 years. *Perceptual & Motor Skills*, 112, 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Matarese, V. (2022). Kinds of replicability: Different terms and different functions. *Axiomathes*, 32 (Suppl 2), 647–670. https://doi.org/10.1007/s10516-021-09610-2

McShane, B.B.; Gal, D.; Gelman, A.; Robert, C.; & Tackett, J. L. (2019) Abandon statistical significance. *The American Statistician*, 73 (Suppl 1), 235–245. https://10.1080/00031305.2018.1527253

Maxwell, S.E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London, Series A* (Mathematical and Physical Sciences), 236, 333–380. https://doi.org/10.1098/rsta.1937.0005

Neyman, J. (1961, [1]1950). *First course in probability and statistics*. New York: Holt, Rinehart And Winston.

Neyman, J. (1956). Note on an article by Sir Ronald Fisher. *Journal of the Royal Statistical Society: Series B (Methodological)*, 18(2), 288–294. https://doi.org/10.1111/j.2517-6161.1956.tb00236.x

Neyman, J.; & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, 20A(1/2), 175. https://doi.org/10.2307/2331945

Neyman, J.; & Pearson, E.S. (1933). *On the problem of the most efficient tests of statistical hypotheses*. London: Harrison And Sons, Ltd.

Neyman, J.; & Pearson, E.S. (1967). *Joint statistical papers*. Cambridge: Cambridge University Press.

Nosek, B.A.; & Errington, T.M. (2017). Making sense of replications. *eLife* 6:e23383. http://doi.org/10.7554/eLife.23383

Pashler, H.; & Harris, C. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7, 531–553.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). https://doi.org/10.1126/science.aac4716

Olsson-Collentine, A.; Wicherts, J.M.; & van Assen, M.A.L.M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, 146(10), 922–940. https://doi.org/10.1037/bul0000294

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be

reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302), 157–175. https://doi.org/10.1007/978-1-4612-4380-9_2

Pearson, E.S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society: Series B (Methodological*), 17(2), 204–207. https://doi.org/10.1111/j.2517-6161.1955.tb00194.x

Popper, K.R. (1959). *Logic of discovery. London*: Routledge

Redish, A.D.; Kummerfeld, E.; Morris, R. L.; & Love, A. C. (2018). Opinion: Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences of the United States of America*, 115(20), 5042–5046. https://doi.org/10.1073/pnas.1806370115

Rothstein, H.R.; Sutton, A. J.; & Borenstein, M. (eds.) (2005). *Publication bias in meta-analysis*. Chichester: John Wiley & Sons, Ltd.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rossi, J.S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, 58, 646–656. https://doi.org/10.1037//0022-006x.58.5.646

Rubin, M. (2019). What type of Type I error? Contrasting the Neyman–Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*, *198(6)*, 5809–5834. https://doi.org/10.1007/s11229-019-02433-0

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1), 1–6. https://doi.org/10.1086/287231

Schauer, J.M.; & Hedges, L.V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, 146(8), 701–719. https://doi.org/10.1037/bul0000232

Sedlmeier, P.; & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Serra-Garcia, M.; & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science advances*, 7(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Stanley, T.D.; Jarrell, S.B.; & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician*, 64(1), 70–77. http://dx.doi.org/10.1198/tast.2009.08205

Stanley T.D.; Carter, E.C.; & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin,* 144(12), 1325–1346. http://dx.doi.org/10.1037/bul0000169

Schäfer, Th.; & Schwarz, M.A. (2019). The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, 10, 813. https://doi.org/10.3389/fpsyg.2019.00813

Szucs, D.; & Ioannidis, J.P.A. (2017a). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience*, 11, 390. https://doi.org/10.3389/fnhum.2017.00390

Szucs, D.; & Ioannidis, J.P.A. (2017b). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PloS Biology, 15(3),* e2000797. https://doi:10.1371/journal.pbio.2000797

Thorn, F. S.; Fidler, F.; & Dudgeon, P. (2019). The Statistical Power of Psychology Research: A Systematic Review and Meta-analysis. *OSF Preprint*. https://doi.org/10.17605/OSF.IO/H8U9W

Trafimow, D.; Amrhein, V.; Areshenkoff, C.N.; et al. (2018). *Manipulating the alpha level cannot cure significance testing*. *Frontiers in Psychology*, 9, 699. https://doi.org/10.3389/fpsyg.2018.00699

Trafimow, D.; & Earp, B.D. (2017). Null hypothesis significance testing and Type I error: The domain problem. *New Ideas in Psychology*, 45, 19–27.

van Dongen, N.; & Sikorski, M. (2021). Objectivity for the research worker. *European Journal for Philosophy of Science*, 11, 93. https://doi.org/10.1007/s13194-021-00400-6

Wagenmakers, E.-J.; Wetzels, R.; Borsboom, D.; & van der Maas, H.L.J. (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), 426–432. https://doi.org/10.1037/a0022790

Wasserstein, R.L; & Lazar, N.A. (2016). The APA's statement on p-values: context, process, and purpose. *American Statistician*, 70, 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wetzels, R.; Matzke, D.; Lee, M.D.; Rouder, J.N.; Iverson, G.J.; & Wagenmakers, E.J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298. https://doi.org/10.1177/1745691611406923

Witte, E.H.; & Zenker, F. (2017). From discovery to justification: Outline of an ideal research program in empirical psychology. *Frontiers in Psychology*, *8*, 1847. https://doi.org/10.3389/fpsyg.2017.01847

Witte, E.H.; Stanciu, A.; & Zenker, F. (2022). Predicted as observed? How to identify empirically adequate theoretical constructs. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.980261