Cohen's convention and the body of knowledge in behavioral science

[1] Bogazici University, Department of Philosophy, 34342, Bebek, Istanbul; ORCID: 0000-0002-3014-6532

[2] Nankai University, College of Philosophy, Tianjin, P.R. China, ORCID: 0000-0001-7173-7964; [*]corresponding author: fzenker@gmail.com

Abstract: In the context of discovery-oriented hypothesis testing research, behavioral scientists widely accept a convention for false positive (α) and false negative error rates (β) proposed by Jacob Cohen, who deemed the *general relative seriousness* of the antecedently accepted $\alpha = 0.05$ to be matched by $\beta = 0.20$. Cohen's convention not only ignores contexts of hypothesis testing where the more serious error is the β-error. Cohen's convention also implies for discovery-oriented hypothesis testing research that a statistically significant observed effect is four times more probable to be a *mistaken discovery* than for a statistically significant true observed effect to be independently replicable. In the long run, Cohen's convention thus is epistemically harmful to the development of a progressive science of human behavior, making its acceptance crucial in explaining the replication crisis in behavioral science. The balance between α- and β-errors generally ought to be struck using both epistemic and practical considerations. Yet epistemic considerations alone imply that making a genuine contribution to the body of knowledge in behavioral science requires error rates that are not only small but also symmetrical.

Keywords: α- and β-error rates; false positive and false negative test results; inductive risk; null hypothesis significance testing; type I and type II error; utility

Cohen's convention and the body of knowledge in behavioral science

Aran Arslan[1] and Frank Zenker[2*]

[1] Department of Philosophy, Bogazici University, 34342, Bebek, Istanbul; ORCID: 0000-0002-3014-6532

[2] Nankai University, College of Philosophy, Tianjin, P.R. China, ORCID: 0000-0001-7173-7964; [*]corresponding author: fzenker@gmail.com

1. Introduction

As a review of the best behavioral science journals would show, researchers engaged in discovery-oriented hypothesis testing have conventionally accepted the false positive and false negative error rates proposed, in 1965, by Jacob Cohen. In 1988, Cohen argued as follows for this convention:

> "It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value [the $(1-\beta)$-error rate], the value .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, 1988, 56; *italics added*)

In previous work, Cohen (1970) had specified the focal notion of 'general relative seriousness' as the *costliness* of errors.

> "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 $\alpha = .05$ convention, it suggests the stance that Type I errors are about four times as "costly" as Type II errors, i.e., $\beta / \alpha = .20 / .05 = 4$." (Cohen, 1970, 825)

In 1992, Cohen would further specify 'costly' as the *cost of the sample collection process*. As before, he held that "a materially smaller value than [β =] .80 would incur too great a risk of a Type II error." Additionally, he now claimed that a "materially larger value [than β = .20] would result in a demand for [the sample size] *n* that is likely to exceed the investigator's resources" (Cohen, 1992, 156).

Our primary aim is to clarify the conditions under which the 'general relative seriousness of errors' renders Cohen's convention reasonable, respectively when it becomes unreasonable. Our secondary aim is to show that the wide acceptance of Cohen's convention is crucial in explaining the replication crisis in behavioral science (Ioannidis, 2005; Open Science Collaboration, 2015).

Beginning with a review of Cohen's convention (Sect. 2), we interpret the "general relative seriousness of errors' along two dimensions. Preferentially minimizing the inductive risk of an α-error rate in discovery-oriented hypothesis testing researcher can be *epistemically* warranted, whereas the (negative) utility associated with a false hypothesis test result can *practically* warrant to preferentially minimize the β-error rate (Sect. 3). An individual researcher's seemingly reasonable preference for asymmetrical error rates in discovery-oriented research, however, is in collectively harmful, for the probability that a genuine discovery will *replicate* is in the long run four times lower than the probability of a making a *mistaken* discovery (Sect. 4). A genuine contribution to the body of knowledge in behavioral science, therefore, requires error rates that are not only small but also symmetrical (Sect. 5).

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

2.1 Discovery-oriented hypothesis testing research

In behavioral science as elsewhere, an evidence-based decision to maintain or reject a hypothesis in view of measurement scores that are sampled from some population of interest is subject to various kinds of error. The measurement scores associated with behavioral responses can hence be related only *indirectly* to empirical hypotheses by an intermediate use of statistical inference procedures as tools. Use of these procedures thus implies an ontological transformed of raw measurement scores ("observations") into probability density distributions ("observed data").

Although the Bayesian hypothesis testing paradigm (Fienberg, 2016) does increasingly gain in prominence, the default statistical paradigm in behavioral

science remains *null hypothesis significance testing* (NHST) (Gigerenzer, 1987; 2004; Morrison & Henkel, 1970). In NHST, researchers compare data (D) to a *null hypothesis* (H_0) stating a zero-correlation or zero effect between focal variables. Additionally, D can be compared to an *alternative hypothesis* (H_1) stating an effect or correlation that is non-zero (directional hypothesis) or of some definite strength (point hypothesis). In both cases, researchers who seek to contribute a *discovery* to scientific knowledge rely on a criterion for *rejecting* the H_0 that is given as the *p*-value of data in view of the H_0, $p(D, H_0)$, falling below the statistical significance level ($p = 0.05$).

A progressive science of human behavior must ultimately aim at fallible knowledge of the *likelihood* (Edwards, Lindman & Savage, 1963; Edwards, 1972) of a theoretically predicted hypothesis given new data, $L(H_x \mid D)$, where $x = 0,1$. Notice that the likelihood ratio $LR_{H1/H0}$ is *equivalent* to the Bayes factor, $BF_{H1/H0}$, in case the prior probabilities of two rivaling hypotheses are uninformative and non-distributed (Witte & Zenker, 2017; Krefeld-Schwalb, Witte & Zenker, 2018). All that NHST can provide, however, is the *probability* of already observed data in view of H_x, $P(D, H_x)$. This probability provides an evidence-based warrant to *reject* H_x if observed data are inconsistent with it, respectively to *maintain* H_x if observed data are consistent with it. Whereas an evidence-based warrant to *accept* H_x because new data confirm it does require the *likelihood* that NHST precisely cannot provide.

The decision (not) to reject H_x is associated with two types of errors. A *true* hypothesis under test may be mistakenly rejected or a *false* hypothesis under test may be mistakenly maintained. Rejecting the H_0 thus entails maintaining the H_1, and vice versa. In the Neyman-Pearson version of NHST (Neyman & Pearson, 1967)—itself an advancement of Fisher's (1956) version, which considers only $P(D, H_0)$—the *long-run* chance of rejecting a true H_0 is called the α-error rate (false positive error; Type I error) and the long run chance of maintaining a false H_0 is called the β-error rate (false negative error; Type II error). Depending on whether the H_0 or the H_1 is consistent with data, then, a hypothesis test has four possible outcomes (Table 1).

Table 1

Confusion matrix for the possible outcomes of a hypothesis test

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

2.2 Cohen's convention

Forwarded against the background of the Neyman-Pearson version of NHST, Cohen's convention can be stated as follows:

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

Consequently:

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

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, 56). Failure to find something can be explained either as a *failed* discovery (not finding something that *is* there) or

as a *null result* (there being nothing to be found). By contrast, finding something that is *not* there is singularly explainable as a *mistaken* discovery.

Although Cohen is silent on how the relative seriousness of α- and β-errors ought to be evaluated, he states that α-errors are about *four* times as serious as β-errors (ibid.). His specific proposal thus is that researchers who antecedently accept $\alpha = 0.05$ default on $\beta = 0.20$. Advocating the conventional acceptance of an error rate ratio of $\alpha / \beta = 0.05 / 0.20 = 1 / 4$, therefore, is to advocate that the probability of mistakenly rejecting a *true* H_0-hypothesis (α-error) be set to *one-fourth* of the probability of mistakenly maintaining a *false* H_0-hypothesis (β-error).

Cohen presumably recognized that the relative seriousness of a β-error may exceed that of an α-error in particular contexts (see Sect. 4.) Whether he found these contexts negligible one cannot know because other than by eventually pointing to resource restrictions, he does not explain what "other basis for setting the desired power value" (Cohen, 1988, 56) an investigator *ought* to have. Of course, his proposal to default on $\alpha / \beta = 0.05 / 0.20$ must somehow relate to the 'general relative seriousness' of errors. But this notion he does not fully explain either.

2.3 The general relative seriousness of errors

To develop this explanation, we can metaphorically refer to the literature where a discovery is published as 'the body of scientific knowledge'. A major assumption in NHST is that this body cannot be harmed (nor improved) if researchers *maintain* the H_0 in response to obtaining a statistically *insignificant* hypothesis test result, $P(H_0, D) > \alpha$. The very same assumption also explains the praxis of selectively publishing statistically *significant* hypothesis test, i.e., a publication bias *pro* discoveries. Regardless of whether the H_0 is maintained given a null result or a missed discovery (β-error), therefore, statistically insignificant test results are unlikely to enter the body of scientific knowledge, thus making no difference to it.[1]

[1] Although the praxis of selective publishing did recently begin to change, that statistically insignificant hypothesis test results hence remain hard to access (*file drawer problem*; Rosenthal, 1979) entails that researchers cannot easily correct the meta-analytical estimated population effect sizes that discovery-based meta-analyses consequently tend to *over*estimate (Rothstein, Sutton & Borenstein, 2005). Selective

In the context of discovery-oriented research, maintaining the H_0 can thus be compared to a safe bet. Whereas if a false positive test result (α-error) is published, then what enters the body of scientific knowledge is a *mistaken discovery*, one that must be thought to "linger" until subsequent research corrects it. The seriousness of a mistaken discovery entering the body of scientific knowledge thus points to the *epistemic* risk of being misled by a falsity. In discovery-oriented hypothesis testing research, therefore, a mistaken discovery entering the body of scientific knowledge is a *more serious* error than a missed discovery (β-error).

Cohen, we may assume, would hence have reasoned that researchers are *epistemically justified* to prefer maintaining a false H_0 over rejecting a true H_1. At the same time, Cohen would have reasoned that, if $\beta > 0.20$, then *too many* missed discoveries do in the long run *fail* to contribute to the body of scientific knowledge. Thus, like the seriousness of including in the body of scientific knowledge what is *not* there (α-error) being justified epistemically, so would $\beta = 0.20$, namely by the seriousness of otherwise remaining ignorant of what *is* there.

Both $\alpha = 0.05$ and $\beta = 0.20$ thus appear to be justified epistemically. But Cohen's additional reason—that a "materially larger value [than $\beta = .20$] would result in a demand for [the sample size] n that is likely to exceed the investigator's resources" (Cohen, 1992, 156)—may suggest that Cohen's justification specifically for $\beta = 0.20$ instead grounds in the *practical* consideration that resource restrictions limit the sample size researchers can collect. The justification for $\alpha = 0.05$ would thus remain epistemic, while the justification $\beta = 0.20$ would be practical.

---

publishing ultimately reflects Popper's (1959) *falsification principle*: a hypothesis can be falsified, but not verified. Statistically insignificant hypothesis test results thus are *uninformative* in NHST because researchers can maintain the H_0, but they cannot interpret data as verifying the H_0 (see Sect. 2.1). Arguably the best countermeasure against the negative effects of selective publishing is to combine *results blind manuscript evaluation* (RBME), whereby publication decisions become independent of statistical significance (Locascio, 2019; Berlin & Ghersi, 2005; Chambers 2013), with a *likelihood ratio* (LR) hypothesis test that leaves statistically insignificant test results informative towards verifying the H_0 (Witte & Zenker, 2017; Krefeld-Schwalb, Witte & Zenker, 2018).

Yet had Cohen considered $\beta = 0.20$ epistemically unjustifiable in discovery-oriented research, then an *epistemic* warrant would be lacking to consider $\beta = 0.20$ a *reasonable* compromise between the epistemic goal of discovery-oriented research (i.e., making genuine discoveries while avoiding mistaken and missed discoveries) and using the limited resources that researchers have towards this goal. Resource restrictions would thus not only dominate the epistemic goal of discovery-oriented research but *undercut* it. By Cohen's standards, however, that is absurd. For Cohen, the *primary* sufficient reason for 0.20 as an *upper* β-error-bound, therefore, must be the epistemic consideration of limiting the proportion of *missed discoveries* that fail to enter the body of scientific knowledge. Hence, invoking resource restrictions to justify 0.20 also as a *lower* β-error-bound merely provides a *supererogatory* reason (i.e., an additional sufficient reason).

This, we claim, is the *only* reasonable reconstruction of Cohen's justificatory structure. Alternative reconstructions simply cannot account for all considerations that Cohen brings to bear. Of course, presenting this reconstruction is distinct from claiming that Cohen "got it right." Indeed, the following sections argue that Cohen's convention is not only unreasonable in *other* contexts of hypothesis testing but also *long run* unreasonable in the context for which Cohen had proposed it.

3. The shape of knowledge in behavioral science

3.1 The replicability of observed effects

What has rightly raised doubt about the application of NHST in behavioral science is widely referred to as the *replication* or *confidence crisis*. Its statistical mark is that a sizable proportion of published NHST-based studies, the reported effects of which are non-replicable, were observed under low *statistical test power* (Krefeld-Schwalb, Witte, Zenker, 2018; Szucs et al., 2017a; van Dongen et al., 2021; Wagenmakers et al., 2011). Conceptually equivalent to the $(1 - \beta)$-error rate, statistical test power is a function of the sample size (N), the effect size ($d = [(m_1 - m_0) / s]$), and the α-error rate. Statistical test power largely determines the replication probability of a *true* observed effect—largely because replication studies are subject to a regression effect, the size of which correlates inversely with sample size (Fiedler & Prager, 2018). Replication studies, therefore, are likely to observe a smaller effect than an original study.

Other things equal, the $(1 - \beta)$-error rate increases with N (see Table 2 and Fig. 1). As the β-error rate decreases, therefore, the cost of the sample collection process increases.

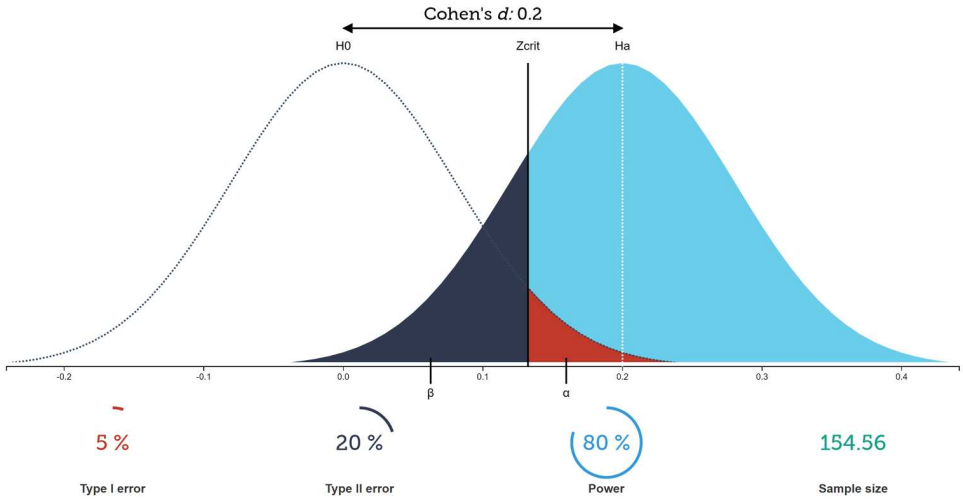
Table 2

The total minimum sample size for both groups (e.g., experimental and control group) in a one-sided t-test as a function of test-power $(1 - \beta)$ and effect size $d = [(m_1 - m_0) / s]$, given $\alpha = 0.05$.

	d			
$(1 - \beta)$	0.01	0.20	0.50	0.80
0.40	38726	97	15	6
0.50	54111	135	22	8
0.80	123651	309	49	19
0.95	216443	541	87	34

Figure 1

The overlap of the probability density distributions for H_0 and H_1 per group for a one-sided t-test, given α=0.05, $(1 - \beta) = 0.80$, and $d = 0.20$.
(*CC-BY license, https://rpsychologist.com/d3/nhst/*)



The best explanation for why published NHST-based studies in behavioral science report effects that are observed under low statistical test power is that

samples are typically too small (Cohen, 1962; 1992; Maxwell, 2004; Rossi, 1990; Sedlmeier & Gigerenzer, 1989; Szucs & Ioannidis, 2017b). For instance, the median total sample size for published studies in psychology is estimated as $N = 40$ (Marszalek et al, 2011; Wetzels et al., 2011; see Bakker, van Dijk & Wicherts, 2012). What contributes to low statistical test power is that "most studies involve tests of multiple hypotheses, [thus] creating a gap between the power for any single test and the power for the collection of tests," wherefore—despite underpowered single tests—"the probability of rejecting at least one hypothesis in the collection of tests will clearly exceed the probability that any specific hypothesis is rejected" (Maxwell, 2004, 148). Low statistical test power, therefore, also results from applying a questionable research practice that increases the probability of observing a *publishable* test result (see our note 1).

Cohen had demonstrated awareness that published effects in behavioral science are typically underpowered already in 1962. He then estimated the *average* statistical test-power in behavioral science as $(1 - \beta) = 0.18$ (Cohen, 1962). Some 30 years later, this estimate had not significantly improved (Cohen, 1992; Maxwell, 2004). Although Cohen thus recognized the need to collect *yet larger* samples, his proposed value $(1 - \beta) = 0.80$ would at least significantly improve over $(1 - \beta) = 0.18$. Because of the minimum sample size (Table 1), indeed, a yet smaller value than $(1 - \beta) = 0.80$ may have appeared unachievable given that published effects in behavioral science are typically *small* ($d = 0.20$). (We return to this in Sect. 5).

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

That samples are typically too small holds both for research in the Neyman-Pearson version of NHST—for which Cohen's convention declares it to be more relevant to control the α-error rate than the β-error rate—as well as for research in the Fisher version of NHST, which recognizes only the p-value as relevant, respectively the associated α-error rate.[2] In critiquing this asymmetrical evaluation of the relevance

[2] Originating in the Fisher version of NHST, the p-value states the probability of observing actual or more extreme data on the assumption that the H_0 is true. Whereas the α-error rate originating in Neyman-Pearson test-theory states the long run probability of mistakenly rejecting the H_0 (false positive). Although differences

of errors, some authors have even proposed to *abandon* NHST (Lakens et al., 2018; McShane et al., 2019; Trafimow et al., 2019). A part of the motivating reason is that researchers regularly misinterpret a hypothesis test where $P(D, H_0) < p = 0.05$ (or $\alpha = 0.05$) as implying a probability > 95% that the H_1 is *true*. Known as the *p-value fallacy* or the *prosecutor's fallacy*, this inference amounts to an unwarranted transition from a probability to a likelihood (see Sect. 2.1), thus reflecting "the mistaken idea that a single number [e.g., $p = 0.05$] can capture both the long run outcomes of a scientific study and the evidential meaning of a single result" (Goodman, 1999, 995; see Cohen, 1994, 997).

Avoiding this fallacy requires no more (nor less) than interpreting the observed *p*-value as stating the probability of observing an effect size equal to, or more extreme than, the observed effect size given the H_0 is true. This interpretation, of course, is easily available. Abandoning NHST, therefore, appears overly extreme. For rather than NHST or the *p*-value being intrinsically faulty (Gómez-de-Mariscal et at, 2021), the main fault lies with their well-documented misapplication.

Equally well-documented is the problem of equating statistical significance with $p = 0.05$, as well the related problem of transitioning without warrant from 'statistical significance' to 'scientific importance'. Recently, a widely cited statement by the *American Statistical Association* (Wasserstein, 2016) once again warned against interpreting the *p*-value as a critical measure of evidence (see Halsey et al., 2015). Indeed, already Fisher (1925), who originally proposed $p = 0.05$ as a conventional error rate (Hubbard, 2016; Kennedy-Shaffer, 2019), did merely offer a *convenience* justification for this specific value.

---

in origin, use, and interpretation must keep from equating the *p*-value and the α-error rate, these differences lack practical bearing when estimating the probability of a *mistaken* decision to maintain or reject the H_0. If the decision criterion is the *p*-value, then the probability estimate is based on data (objective interpretation), whereas if the decision criterion is the α-error rate, then probability estimate is based on a researcher's expected error rate (subjective interpretation). An evidence-based decision nevertheless requires that the subjectively expected α-error rate be set to a value *at least as large* as the objective *p*-value. The conceptual differences between the *p*-value and the α-error rate thus remain "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, 47; *notation adapted*)

With $p = 0.05$ stating roughly the probability that a mean observed effect falls more than two standard deviations away from the mean of a normally distributed random variable, Fisher thus surmised that "we shall *not often be astray* if we draw a conventional line at .05, and consider that higher values of χ^2 [chi-square] indicate a real [rather than a mistaken] discrepancy" (Fisher, 1925, p. 79; *italics added*). (Pearson's (1900) chi-squared test determines the strength of association between two variables in a contingency table).

As the statistical significance threshold for rejecting the H_0, then, already $p = 0.05$ could be accepted merely conventionally, because a justification to prefer $p = 0.05$ to any other p-value was lacking already in Fisher's (1925) statistical inference system, on which NHST is based. Add to this that when, in 1885, Edgeworth coined the term 'statistical significance', he merely intended "to have 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, 2; *italics added*). This makes it more understandable why despite the frequent misinterpretations of the p-value, the conventional acceptance of its probability-based definition remains convenient (see Kennedy-Shaffer, 2019, 84).

3.2 Conventions by convention?

Although both NHST and alternative statistical inference systems rightly count as well-established today, behavioral scientists seem to apply these systems for the most part *conventionally*. Conventions, however, cannot justify their own application, a truism that Cohen (1965) and Neyman and Pearson (1933) fully

acknowledged. To better understand why behavioral scientists conventionally adopted $\alpha = 0.05$ and $\beta = 0.20$, the following passage by Neyman and Pearson (1933) is 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, 296; *italics added*)

Although striking the right balance between both error rates was thus left to researchers themselves, Neyman and Pearson likewise advocated an α-error rate of $\alpha = 0.05$ so that "in the long run of experience, we shall not too often be wrong" (Neyman & Pearson, 1933, 291). Since Neyman (1950, 262) had already suggested that α-errors are more serious than β-errors, all that Cohen added was the point-specific ratio $\alpha / \beta = 1 / 4$. But not until Cohen (1992) cited resource restrictions did his justification exceed the claim that this ratio ought to be evaluated according to standards the scientific community already accepts conventionally. Precisely why Cohen (1958; 1965) proposed the "arbitrary but reasonable value" (Cohen, 1988, 56) of $\beta = 0.20$ thus remained vague, or so the following quote shows:

> "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, 1958, rev. ed., 1965, 98f.; *italics added*)

Indeed, Cohen had broadly *avoided* stating how 'general relative seriousness' should be interpreted. Although he indicates—notice how the term 'serious' simply reappears—the lack of *serious* reasons to oppose conventions other than $\alpha / \beta = 1 / 4$, he simply put one convention on top of another. Since scientists already accepted $\alpha = 0.05$ conventionally, Cohen could thus recommend the *dependent* convention of $\beta = 0.20$. More recent scholarship reiterates the same idea. "[I]n the internal dealings of science," for instance, "errors of Type I [α-error] are in general regarded as *more problematic* than those of Type II [β-error]" because "those who claim the existence of an as yet unproven phenomenon have the burden of proof" (Hansson, 2018, 7; *italics added*).

Cohen's (1992) justification of why science gives *higher priority* to avoiding an α-error than a β-error, and specifically his justification for $\beta = 0.20$, eventually cited that resource restrictions limit the sample size that a researcher can collect (see Sect. 2.3). The general relative seriousness of errors thus *seemed* to be justified by a *practical* consideration. But we saw that this can merely provide a supererogatory reason for $\beta = 0.20$. The primary sufficient reason for $\beta = 0.20$, and thus for keeping "the relation between the two risks [α- and a β-error] within reasonable bounds" (Cohen, 1965, 98f), was the *epistemic* consideration of limiting the proportion of *missed discoveries* not entering the body of scientific knowledge.

With Cohen, then, if the probability of avoiding α- and β-errors ought to mirror their general relative seriousness, then the more serious an error is, the less likely a reasonable researcher would want its occurrence to be. Other things equal, therefore, if an α-error takes one-fourth of the numerical value of a β-error—because rejecting a *true* H_0 ought to be four times less probable than not rejecting a *false* H_0—then a true H_0 would in the long run be mistakenly rejected four times less often than a false H_1 would be mistakenly accepted.

But is an α-error in fact more serious than a β-error?

4. Two evaluative dimensions

4.1 Epistemic and non-epistemic values

By the mid-20th-century, the philosophy of science debate on hypothesis testing, as well as the concurrent developments in statistics and probability theory, had not only demonstrated the availability of rigorous statistical inference systems that express the degree of confidence in a scientific hypothesis under test (Andersen & Hepburn, 2016, 25) but had also suggested that understanding 'hypothesis testing' as a *decision* between possible actions should go along with acknowledging a *value* component (ibid.). Yet the extent to which a correct decision on maintaining or rejecting a hypothesis is "driven" by this value component is no less controversial than the idea that the decision itself requires not only *epistemic* but also *non-epistemic* values.

Associated with the *truth-likeness* of a hypothesis, epistemic values are commonly exemplified by truth, simplicity, explanatory power, or predictive accuracy (e.g., Kuhn, 1962; 1977). Whereas non-epistemic values, associated with the *utility* of a scientific study's results, are commonly exemplified by moral, legal,

or social values of relevance for public policy making. Hence, if the evaluative criterion for a hypothesis under test is the general relative seriousness of errors, then the balance between α- and β-errors ought to be struck in ways that are sensitive to the epistemic and practical consequences of these errors.

> "Many controversies on risk assessment concern the balance between risks of type I [α-errors] and type II errors [β-errors]. Whereas science gives higher priority to avoiding type I errors than to avoiding type II 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, 7)

Although similar examples where this balance can shift are easy to come by (e.g., regarding legal cases, the environment, or public health), they are *orthogonal* to Cohen's convention, which is reasonable only if failing to discover what is there is less serious for the body of scientific knowledge than mistakenly "discovering" what is not there. Given that making a genuine scientific discovery, as well as avoiding mistaken and missed discoveries, are *epistemic* goals, therefore, accepting Cohen's convention is epistemically warranted because preferring a missed over a mistaken discovery is epistemically warranted.

Yet this preference reverses when the seriousness of errors is evaluated based on the (negative) *utility* of mistakenly maintaining a hypothesis under test. The key contrast thus is that between epistemic and practical considerations of hypothesis testing, a contrast at the heart of a debate 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 the H_0 can be accepted as true', the *truth* of H_0 counts more than its *utility*. On his view, if the available evidence consistent with the H_0 is scant compared to available evidence for an equally plausible alternative hypothesis, a researcher should *reject* the H_0 even if it is true.

Unlike Fisher, who treated significance tests and *p*-values as continuous measures of evidence against the H_0, Neyman and Pearson (Neyman, 1956; Pearson, 1955) primarily addressed whether researchers should *act* as if the H_x were true. Although they acknowledge that an evidence-based decision between rejecting or maintaining the H_0 or the H_1 must be sensitive to the seriousness of both errors, they understood this decision—contra Fisher—to depend on both the available evidence and its utility.

Speculatively, in recommending increased caution in avoiding α- rather than β-errors, Cohen may have recured to utility considerations *implicitly*. For Cohen agrees with Neyman and Pearson that an α-error is more serious than a β-error. And he agrees with Fisher that a hypothesis test seeks to determine a hypothesis's truth (rather than maximizing a decision's utility), as well as that a hypothesis, even if true, ought to be rejected if the supporting evidence is scant relative to undermining evidence. The last point of agreement between Cohen and Fisher may particularly suggest that utility considerations are implicit in Cohen's notion of evidence. His *primary* reason for α / β = 0.05 / 0.20, however, is *epistemic*—namely to limit the proportion of mistaken discoveries entering the body of scientific knowledge. In Cohen's justification, then, utility considerations fail 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 two kinds of risk-related decisions (Hansson, 2018). Based on epistemic considerations, the first kind of decision concerns the identity of risk-related information the body of scientific knowledge should include if individual researchers *publish* the decision reached in discovery-oriented research as a scientific result. In this context, the seriousness (S) of a *mistaken discovery* is justified by potentially being misled by a falsity. Compared to a *failed discovery*, therefore, itself thought to be unlikely to enter that body, a mistaken discovery is a more serious error. In brief, $S(\alpha) > S(\beta)$.

To limit the risk of a mistaken discovery, science offers *epistemic* reasons to endorse strict proof standards (Hansson, 2018). The five-sigma ($5 \times \sigma$) standard endorsed in the current standard model of physics, for instance, corresponds to an α-error rate of $\alpha = 0.00003$. It implies that the H_0 will be rejected, only if an observed mean (H_1) deviates by at least five standard deviations from an expected mean (H_0). A similarly small deviation would hence be expected roughly once in three million

tests if the H_0 is true (Bird, 2018, 17). It thus remains highly unlikely for a mistaken discovery to enter the body of scientific knowledge. In behavioral science or medicine, by contrast, where a hypothesis is normally tested at $\alpha = 0.05$, or a proof standard of $1.96 \times \sigma$ (Bentley, 2021, 2), a mistaken discovery would occur roughly once in 20 tests.

What Cohen's convention seemingly omits to consider is the *second kind of decision*, the seriousness of which is additionally sensitive to the (negative) utility of a missed discovery. A paradigm example from a medical context is to mistakenly diagnose a person as diseased, although they are healthy. Compared to the first decision (*mistaken discovery*), the seriousness of a mistaken diagnosis (*missed discovery*) implies a change in the functional role of risk-related information. We can therefore have it that $S(\beta) > S(\alpha)$, namely if a β-error is associated with a larger (negative) utility than an α-error. Since this suggests it is more important to control for the β- than for the α-error rate, $S(\beta) > S(\alpha)$ would imply $\alpha > \beta$.

For instance, assume the use of a sufficiently reliable diagnostic test[3] to test person *P* for a potentially fatal contagious infection, *I*. Cohen's convention would state that the seriousness of *P not* being in condition *I* given the test says *P is* in condition *I* (α-error), exceeds that of *P being* in condition *I* given the test says *P is not* in *I* (β-error). But this cannot be right. Someone who is mistakenly diagnosed as non-infected presents a risk of spreading *I* that someone mistakenly diagnosed as

[3] To use a current example, the accuracy of the *Reverse-Transcription Polymerase Chain Reaction* (RT-PCR) test, a common diagnostic test for SARS-COVID-19, varies with the general laboratory conditions and the kind of polymerase used. Recent studies estimate an RT-PCR test's false positive error rate under *real* (versus test validation) conditions as $0.02 < \alpha < 0.09$ (Andrew, 2020) and its false negative error rate as $0.01 < \beta < 0.30$ (Arevalo-Rodrigez et al., 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 must expect $\beta = 0.21$, while $0.167 < \alpha < 0.29$ (ibid.). The staggering range of the β-error implies that a *single* RT-PCR test "cannot be used to 'clear' people as being non-infected" (Bentley, 2021, 9), wherefore two (or more) *independent* RT-PCR tests are commonly administered because the error rates of *n* independent tests are a multiplicative combination of the error rates of *n* individual tests.

infected cannot present, thus becoming harmful both to themselves and the public. In the false positive case (*mistaken discovery*), *P* may needlessly quarantine, perhaps becoming bored—no doubt a mild negative consequence. Whereas in the false negative case (*missed discovery*), as the risk of spreading the infection endangers public health, the negative consequences here may even be tragic. The relative seriousness of both errors, therefore, favors avoiding a missed discovery (β-error).

Both epistemic and practical considerations thus must be applied to evaluate the general relative seriousness of errors in contexts of public health. Similar contexts point back at the question considered by Laplace of how many votes in a court of judges are needed for a conviction (see Sect. 3.2). If the seriousness of a mistaken verdict, test result, or diagnosis depends on utility considerations, as Neyman and Pearson acknowledge, then the negative utility of an error may vary widely—indeed between boredom and death. To strike a balance between both errors-rates, therefore, adopting a convention that ignores utility considerations, as Cohen's convention does, is poor advice.

4.3 The argument from inductive risk

This point generalizes. As utility considerations come into focus, *any* decision (not) to reject a hypothesis under test had better consider the value-laden character of science, because non-epistemic considerations can be at least as important as epistemic ones (Diekmann & Peterson, 2013). Particularly in the construction and selection of scientific models or in engineering, if the error that is preferentially minimized has public policy implications, the review and evaluation of scientific knowledge cannot be left to experts alone (Lemons et al., 1997). Instead, the public should have a say on, and arguably some control over, how scientific knowledge is used and produced (ibid., 234).

Yet the idea of *value-free science* states that non-epistemic considerations should be absent from the justification of scientific knowledge (Betz, 2016). This ideal has been variously objected to based on the *argument from inductive risk* (e.g., Douglas, 2009; John, 2016; Rudner, 1953). Although this argument can hardly overthrow the value-free ideal of science, it provides compelling reasons that the balance between α- and β-error rates ought to be struck based on epistemic considerations (e.g., a hypothesis's truth-likeness) and non-epistemic considerations (e.g., the utility of maintaining or rejecting H_x) that affect the seriousness of errors.

The argument from inductive risk (John, 2016, 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.
5. *Therefore*, scientific inference must appeal to non-epistemic considerations.

Of course, scientific inference *primarily* requires an epistemic standard. The more stringent this epistemic standard is—i.e., the more evidence of a specific kind is required—the less likely scientists are to maintain a falsehood or to reject a truth. Since a very stringent epistemic standard would remain appropriate even if it were sensitive to the negative utility of error, practical considerations may appear to be *derivative* of epistemic considerations. Whereas if the expected disutility of error cannot be evaluated without appealing *directly* to non-epistemic considerations—as the argument from inductive risk claims—then the indispensability of non-epistemic considerations in scientific inference would become more plausible.

Whereas the argument's first three premises are widely accepted today, the fourth premise may raise suspicion. Does determining the trade-off between both types of error *indispensably* require that the consequences of errors are grounded by appealing directly to non-epistemic considerations? In criticizing that an inclusion of non-epistemic considerations would be *neutral* for scientific inference, for instance, Hudson (2022, 211) rather takes non-epistemic considerations to necessarily lead systematically *away* from truth, predictive accuracy, even logical consistency, and thus to amount to a scientific *bias*. Douglas & Elliott (2022, 202), however, rightly object that *bias* and *value-ladenness* are distinct concepts. A cognitive bias, after all, need not entail value-ladenness, nor vice versa.

Moreover, as the example of testing for a contagious disease had shown, to explain why a β-error is in some contexts of inductive risk more serious than an α-error may indispensably require a direct appeal to non-epistemic considerations. Of course, acknowledging as much cannot vindicate the *general* form of the argument from inductive risk. But that the relevant inductive risk cannot be explained without

acknowledging that scientific inference is sensitive to non-epistemic considerations *in these contexts* nevertheless justifies their local indispensability. And no more is needed.

The one non-epistemic consideration that Cohen's convention recognizes as a supererogatory reason for $\beta = 0.20$ in the context of discovery-oriented hypothesis testing research is the cost of the sample collection process. Whereas his *primary* reason for $\beta = 0.20$ in this context is the consideration of limiting to an epistemically acceptable level the proportion of mistaken or missed discoveries entering or not entering the body of knowledge. As the example of testing for a contagious disease had shown, however, Cohen's justification has obvious faults in other contexts. This is a sufficient reason to reject Cohen's convention *outside* the context of discovery-oriented hypothesis testing research.

5. Long-run epistemic consequences

5.1 Individual vs collective reasonableness

A sufficient reason to reject Cohen's convention *within* the context of discovery-oriented hypothesis testing research is that it had "set the tone" for *asymmetrical* error rates in behavioral science. Although $\alpha / \beta = 0.05 / 0.20$ can be reasonable for an *individual* researcher's contribution to scientific knowledge, asymmetrical error rates entail the long-run failure at the *collective level* to leverage the best systems of statistical inference to develop a progressive science of human behavior. This does not only go a long way toward explaining the replication crisis in behavioral science experiences; it also suggests that this crisis is self-inflicted and avoidable.

This explanation counts because, as several systematic attempts at replicating a key selection of originally observed effects have broadly failed (e.g., Many Labs Projects 1-5, see Ebersorle et al., 2020), *most* published effects in behavioral science should be thought of as non-replicable. We saw that behavioral science journals typically published an observed effect only if it met the conventional significance level $P(\mathrm{D}, \mathrm{H}_0) < \alpha = 0.05$ (Sect. 2.3). We also saw that the replication probability of a *true* observed effect is largely determined by a study's $(1 - \beta)$-error rate, i.e., statistical test power, itself a function of the α-error rate, the observed effect size ($d = [(m_1 - m_0) / s]$), and the study's sample size (N) (Sect. 3.1). Recall that other things being equal, the $(1 - \beta)$-error increases as N increases.

It should now be easy to see that, if N and d are constant, then asymmetrical error rates of the form $\beta > \alpha$ imply that the significance level will never match the statistical test power level. Per Cohen's convention, then, if a statistically significant *true* effect was originally observed under $(1 - \beta) = 0.80$, then the probability that an independent replication study will observe a *very similar* effect size is only 80%. In the long run, therefore, roughly 20 out of 100 replication studies would be expected to fail. This 80 : 20 proportion Cohen presumably found acceptable.

But in psychology, for instance, the largest behavioral science today, *median* observed statistical test power for published studies is estimated to be a mere $(1 - \beta) = 0.35$ (Bakker et al., 2012)—massively undermining Cohen's convention (see Christopher, 2019; Open Science Collaboration, 2015; Stanley et al., 2018). If anything, this is consistent with Cohen's convention having "set the tone" for large β-error rates. At the *collective* level, therefore, the negative consequence for the development of a progressive science of human behavior is that *most* replication attempts in psychology are expected to fail.

More precisely, for statistically significant observed effects published in cognitive neuroscience and psychology journals, median observed statistical power for *small* ($d = 0.20$), *medium* ($d = 0.50$), and *large* effects ($d = 0.80$) is estimated as, respectively, $(1 - \beta) = 0.12, 0.44, 0.73$ (Szucs & Ioannidis, 2017b). *Large* observed effects would thus at least seem to approximate the 80 : 20 proportion of replicable effects entailed by Cohen's convention. Yet the typical pattern of observations in behavioral science is: 'small d, small s' and 'medium-to-large d, large s' (Linden & Hönekopp, 2021; Olsson-Collentine, Wicherts & van Assen, 2020; Schauer & Hedges, 2020). What can be observed with *precision* (read: small s), therefore, are *small* effects alone.

But small effects can be accounted for exclusively by random influences on an empirical setting. This makes small effects poor candidates for the theoretical constructs that a progressive science of human behavior would have to develop. Whereas the large heterogeneity (read: large s) of *medium*-to-*large* effects implies *imprecise* observations, making these effects equally poor candidates for this purpose. And what other purpose would discovery-oriented research in behavioral science ultimately have if not that of *precisely* identifying the *replicable* effect that progressive science of human behavior requires?

A progressive science of human behavior thus entails developing theoretical constructs for replicable effects that have been precisely identified, in turn entailing a small *s*. The most direct way of obtaining a small *s*-value is to collect a large enough sample (*law of large numbers*), in turn increasing the $(1 - \beta)$-error. But if already an original study regularly fails to collect a large enough samples, then a progressive science of human behavior obviously cannot be developed. This speaks against Cohen's convention *within* the context of discovery-oriented research for which it was proposed.

Given that a progressive science of human behavior pivots on observing a very *similar* effect size in a *series* of independent replication studies (Witte, Stanciu, Zenker, 2022), if each of $i = 3$ independent replication studies makes observations under $(1 - \beta) = 0.80$, then the series' statistical test power—given as $(1 - \beta)^i$ (Francis, 2012)—is close to chance ($0.80^3 = 0.51$). Although each independent replication study would thus have to already secure maximal statistical test power, it is more important than the value of $(1 - \beta)^i$ that each study observes a very *similar* effect size. Evaluating whether this is the case requires an *accurate* effect size estimate, and an effect size estimate's accuracy again increases with *N*. The need to collect a sufficiently large *N* thus remains. But this need is precisely what the $\beta = 0.20$-part of Cohen's convention denies.

On exclusively *epistemic* considerations, therefore, low observed statistical test power plausibly provides the major explanatory factor for the replication crisis in behavioral science.

5.2 Reversible vs. irreversible experimental units

A potential objection against this claim states (correctly) that the error rates of Neyman-Pearson test-theory can be unproblematically interpreted as the long-run rate of successful *exact* or *direct* replications. An exact replication seeks to duplicate all aspects of an original study potentially affecting the originally observed effect, whereas a direct replication duplicates merely aspects thought to be theoretically relevant to it. Moreover, the application of Neyman-Pearson test-theory to a series of replication studies requires that the sampling procedure is randomized and that the original study's methodology remains essentially unchanged (Neyman, 1937, 334-335; Neyman & Pearson, 1928, esp. 177, 231, 232).

Under these conditions, the H_0 can be *mistakenly* rejected only because of the random measurement error (reflecting that random factors influence an empirical setting) or the random sampling error (reflecting that a given sample is imperfectly representative of the population). Whereas if the sampling procedure is altered between the *n*-th and the *n* + 1-th study—such that the *n*-th study samples from a different population than the *n* + 1-th study—or if causally relevant aspects of the methodology are changed, then the error rates of Neyman-Pearson test-theory cannot be interpreted as the long-run rate of successful *exact* or *direct* replications.

As changes in methodology or sampling procedure are normal in behavioral science, however, this seems to void interpreting the error rates of Neyman-Pearson test theory as the long-run rate of successful *exact* replications (Rubin, 2019, 202). Failure to replicate an original observed statistically significant true effect would hence be unsurprising. As Rubin (2009) notes for *direct* replications, moreover, the identification of theoretically relevant aspects entails "a theoretical commitment based on the current understanding of the phenomenon under study, reflecting current beliefs about what is needed to [causally] produce a finding" (Nosek & Errington, 2017). But since behavioral science addresses *irreversible* experimental units (e.g., people, social groups, social systems), an irreversible experimental unit that plays a *non-negligible* causal role in producing the original effect would make it *conceptually* impossible to replicate that effect.

For these reasons, Rubin (2019) proposes to resort to the Fisherian sample-specific α-error probability. Instead of being interpreted according to Neyman-Pearson test-theory—i.e., "in relation to a series of samples that could have been randomly drawn from the exact same null population" (Rubin, 2019, 213)—the Fisherian sample-specific α-error probability is interpreted relative to a single, time and location specific sample. As this interpretation breaks with the idea of an exact replication, changes in theoretically relevant aspects would seem unproblematic. But this interpretation also relegates observed effects in behavioral science to the realm of *contingently* repeatable findings. Hence, the argument that failing to replicate observed effects is problematic would be a non-starter.

The kind of replication study that Rubin neglects, however, is a *conceptual* replication. Even if experimental units differ, exact and direct replication studies not only rely on the same *operationalization* as an original study but also identify the same *theoretically relevant* aspects. Conceptual replications, by contrast, seek to

replicate an originally observed effect by identifying *different* theoretically relevant aspects. This allows researchers to sample from different populations, manipulate irreversible experimental units, and use different operationalizations. Compared to an effect that is replicable only under the same operationalization, indeed, one that is replicable under different theoretical frameworks is arguably more likely to be a *true* positive effect (Crandall & Sherman, 2016). Therefore, the observation that given different operationalizations or irreversible units the error rates of Neyman-Pearson test-theory cannot be applied, is ultimately misplaced.

5.3 Does the replication crises matter?

A different question altogether is whether the replication crisis deserves attention. Several scholars argue that the replication crisis is *not* a serious problem (e.g., Redish et al., 2018). Lewandowsky et al. (2020, 3) even claim that—despite such questionable research practices (QRPs) as Hypothesizing After the Results are Known (HARKing), *p*-hacking, or publication bias being possible causes of the replication crisis—in an *idealized* transparent scientific community that abandons QRPs, a low replicability rate *supports* robust and efficient science (ibid., pp. 2f.) because it *reduces* the cost of acquiring scientific knowledge while simultaneously *increases* its efficiency. This idealized community is modeled such that either "all findings are replicated before publication to guard against replication failures" (ibid.) or "individual studies are published and are replicated after publication, but only if they attract the community's interest" (ibid.). Community interest here is equated with the observed citation pattern for a published study.

The claim that a low replicability rate *supports* robust and efficient science is based on simulations of discovery-oriented studies that reveal the cost of generating scientific knowledge to vary considerably between both replication regimes. Compared to leaving it to the scientific community to show interest, the first regime "incurred an additional cost of around ten studies […] [,] represent[ing] ~10% of the total effort the scientific community expended on data collection" (ibid., 4). Although the "analysis of replicability confirms that citations do not predict replicability" (ibid.), it holds regardless of the regime that "the probability of replication of a study *increases* with the number of citations" (ibid., 3; *italics added*). The efficiency of knowledge generation, therefore, is proportional to the number of potentially nonreplicable published studies.

However, not only is Lewandowsky et al.'s (2020) idealized community highly idealized. The *simulated* efficiency of knowledge generation without QRPs is also trumped by how knowledge is *in fact* generated. Independently of the number of citations, after all, compared to a statistically significant effect that is observed in a *well-powered* study ($\alpha = \beta < 0.05$), a statistically significant effect of similar size that is observed in an *underpowered* study is more likely to be a β-error. For corresponding simulation, see Witte, Stanciu, Zenker (2022).

Bird (2018) likewise acknowledges the problematic role of QRPs, as well as the unsound application of statistical methods. Yet he takes the replication crisis to indicate that behavioral science is secure because a high proportion of failed direct replications is the *expected* outcome of high-quality science, if "the field of science in question produces a high proportion of false hypotheses prior to testing" (Bird, 2018, 1). Of course, given the lenient $\alpha = 0.05$, if the base rate of true H_0 hypotheses that are *rejected* is high, then a large proportion of published H_1 hypotheses that are false will "survive" testing (see Sect. 4.2). A good estimate of the actual base rate, however, requires a good estimate of the base rate of *unpublished* studies that *correctly* maintain the H_0. But this estimate, which points back at the file-drawer problem (Rosenthal, 1979), is highly uncertain (see our note 1).

Finally, the recent proposal to decrease the α-error rate rather than the β-error rate (Bartos & Maier, 2019; Benjamin et al, 2018; Bird, 2018; Lakens et al., 2018) ignores that the relative importance of errors varies within and between study areas and researchers, as well as across studies (Trafimow et al., 2018). Thus, "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). But we saw that practical considerations can justify asymmetrical error rates (e.g., $\alpha > \beta$ in contexts of public health). Unless $\alpha = \beta$, therefore, for a progressive science of human behavior even a wide consensus on a very small α-error rate in discovery-oriented hypothesis testing research is mere cosmetics.

5. Conclusion

In proposing that behavioral scientists conventionally default on error rates for false positive (α) and false negative errors (β) that mirror their general relative seriousness, Jacob Cohen had evaluated an α-error to be about four times as serious as a β-error. The antecedently accepted error rate $\alpha = 0.05$ was thus matched with β

= 0.20. Widely accepted in behavioral science today, this convention "set the tone" for accepting *asymmetrical* error rates of the form $\alpha < \beta$.

Cohen's justification for $\alpha / \beta = 0.05 / 0.20$ as a reasonable convention for the context of discovery-oriented research did primarily ground in the *epistemic* consideration of preferentially limiting the proportion of *mistaken discoveries* (as opposed to the proportion of *missed discoveries*) that an individual researcher "contributes" to the body of scientific knowledge. The practical consideration that resource restrictions limit the sample size that a researcher can collect did merely provide a *supererogatory* reason for $\beta = 0.20$. In other contexts of hypothesis testing, however, for instance that of public health, $\alpha / \beta = 0.05 / 0.20$ is an unreasonable convention, because utility considerations decisively show that rejecting a true H_1 (β-error) can be more serious than maintaining a false H_0 hypothesis (α-error). This holds for individuals and social groups alike.

Epistemic considerations can thus suffice to justify favoring a low α-error rate over a low β-error rate in discovery-oriented hypothesis testing research because *individual* researchers here are epistemically justified in preferring a missed to a mistaken discovery. For the development of a progressive science of human behavior, however, asymmetrical error rates of the form $\alpha < \beta$ do necessarily incur long-run negative epistemic consequences at the *collective* level, namely the broad inability to replicate an originally observed true effect. This speaks strongly against Cohen's convention in the context for which he proposed it and makes its wide acceptance crucial in explaining the replication crisis in behavioral science.

Conflict of interest statement

The authors have no conflict of interest to declare.

Author contributions

Both authors drafted the manuscript, which F.Z. edited. Both authors approved the final submitted version.

Acknowledgments

For valuable comments on earlier drafts, we thank audience members at the 2018 Poznan Reasoning Week, 11-15 September 2018 at AMU, Poznan, Poland, and at a talk on 31 May 2022 at Bayreuth University, Germany, as well as the participants of

the POND workshop, 7-8 September 2022 at Bogazici University, Turkey, and Erich H. Witte, Adrian Stanciu, and members of the MTR research group.

Funding

TUBITAK (Project No. 118C257)

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/

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

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

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*, http://dx.doi.org/10.1016/j.jesp.2015.10.002

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. (1988). *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

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

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,* 1–9.

Ebersole, C.R., Mathur, M.B., Baranski, E., Bart-Plange, D.J., Buttrick, N.R., Chartier, C.R., ... & Szecsi, P. (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.

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 (expanded edition, 1992, Johns Hopkins University Press, Baltimore).

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.

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

Fisher, R.A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17(1)*, 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

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.

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

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 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

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.

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

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.

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*(1). 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, sup1*, 346–351. https://doi.org/10.1080/00031305.2018.1505658

Marks, H.M. (2008). *The progress of experiment: science and therapeutic reform in the Unites 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

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

Neyman, J. (1961). *First course in probability and statistics*. 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. (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), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

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.

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

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

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

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

Word count: ca. 9350 (excluding abstract and references)