The case for default point- H_1 -hypotheses in behavioral science: a theory-construction perspective

Frank Zenker^{1*} and Erich H. Witte²

¹ Bogazici University, Department of Philosophy, 34342 Bebek, Istanbul, Turkey;
*corresponding author: frank.zenker@boun.edu.tr; ORCID: 0000-0001-7173-7964

² Institute for Psychology, University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany; witte_e_h@uni-hamburg.de; ORCID: 0000-0002-9446-3211

Abstract: The development of an empirically adequate theoretical construct for a given phenomenon of interest requires an estimate of the population effect size. Arriving at this estimate in evidence-based ways presupposes access to robust findings, defined as statistically significant test-results with high statistical power. In behavioral science, however, even the best journals today typically publish statistically significant test-results with insufficient statistical power, entailing that such findings have insufficient replication probability. Whereas a robust finding formally requires that an empirical study engages with point-specific H_0 - and H_1 -hypotheses, research in behavioral science typically point-specifies only the H_0 , and instead engages with a composite (directional) H_1 . This mismatch renders the prospects for theory-construction poor because the parameter to be theoretically modelled, the population effect size, regularly remains unknown, and so keeps from developing empirically adequate theoretical constructs. Against the background of the research program strategy (RPS), a sophisticated integration of Frequentist and Bayesian statistical inference elements, we claim theoretical progress in behavioral science requires engaging with point- H_1 -hypotheses by default.

Keywords: Bayes factor; Bayesian statistics; Fisher-test; likelihood ratio; meta-analysis; Neyman Pearson test-theory; null-hypothesis significance testing; publication bias; results-blind manuscript evaluation; point-hypothesis; research program strategy; resource pooling; test-power; Wald threshold

The case for default point- H_1 -hypotheses in behavioral science: a theory-construction perspective

Frank Zenker and Erich H. Witte

1. Introduction

2. Hypothesis testing in RPS, NHST, and BHT

2.1 *The research program strategy*

2.1.1 Data-based considerations

2.1.2 Hypothesis-based considerations

2.2 *Frequentist null-hypothesis significance testing (NHST)*

2.2.1 Fisher's NHST-version

2.2.2 Neyman and Pearson's NHST version

2.3 *Bayesian hypothesis testing (BHT)*

2.3.1 Stable test-results

2.3.2 Fairness and firmness of a hypothesis test

3. Composite hypotheses as a barrier to theory-construction

3.1 *Pure induction and the true population effect*

3.2 *Selective publishing and Popper's falsification principle*

3.3 *Induction*

3.4 *The priors*

4. Default point-hypotheses and resource pooling

4.1 *Research programs, default H_1 -hypotheses, and meta-analysis*

4.2 *Underpowered test-results and log-LR addition*

5. Conclusion

Declarations/Acknowledgements

References

1. Introduction

The replication crises in behavioral science implies the widespread failure to repeat published experimental or observational findings by independent efforts. This failure makes it difficult to *trust* observed effect sizes that even the best behavioral science journals report today (Francis, 2014; Ioannidis, 2005; Pashler & Wagenmakers, 2012). The prospects for theory-construction thus are poor because the population effect size, the parameter an *empirically adequate* theoretical construct models, does regularly remain unknown.

The ongoing debate on how to overcome the replication crisis shows many researchers to openly acknowledge a need for an improved statistical inference strategy. Our own point of departure, the *research program strategy* (RPS), grounds statistical inference in Wald's (1947) seemingly forgotten approach to sequential testing (Witte & Zenker, 2017b; Krefeld-Schwalb, Witte & Zenker, 2018). RPS combines select statistical inference elements from Frequentist *null-hypothesis significance testing* (NHST), specifically Neyman Pearson test-theory

(NPTT) (Neyman & Pearson, 1967; Mayo, 1996; Mayo & Spanos, 2011), as well as from *Bayesian hypothesis testing* (BHT) (Jeffrey, 1948; Ly, Verhagen & Wagenmakers, 2016; Morey, Romeijn & Rouder, 2016). A crucial feature is that RPS requires error-rates that are not only small, but also symmetrical ($\alpha=\beta<0.05$), contrasting with Cohen's (1977) widely accepted convention of $\alpha<0.05$, $\beta<0.20$.¹ As computer simulations confirm, error-rates of $\alpha=\beta<0.05$ suffice to obtain statistically significant test-results with high replication probability, that is, *robust* empirical findings (Krefeld-Schwalb, Witte & Zenker, 2018).

As we explain below, a robust finding formally requires that an empirical study engages with H_0 - and H_1 -hypotheses that are specified as mathematical *points*, a requirement that Lakens, Scheel and Isager (2018), for instance, recognize in the comparatively narrower context of NPTT. A typical behavioral science study, by contrast, point-specifies only the H_0 and engages with a *composite* H_1 (aka a *directional* H_1). This mismatch, we claim, keeps from obtaining robust empirical findings, and so explains why the average replication probability of published test-results in behavioral sciences is unacceptably small. This makes it more understandable why few empirically adequate theoretical constructs are available in behavioral science, and why theoretical progress here is hard to discern.²

Arguing from a theory-construction perspective, we here raise the possibly strongest argument for the claim that behavioral science must engage with point- H_1 -hypotheses by default. This argument rests on the well-known formal truth that the probability of successfully replicating an original observed effect size (aka the statistical test-power of an empirical study) is largely determined by the $(1-\beta)$ -error-rate of the data that provide the evidential basis for an original effect size estimate.³ That most published findings in behavioral science are *non-robust* thus

¹ In NPTT, the α -error rate is the long-run proportion of *erroneous* decisions to accept the point- H_0 as true/corroborated. The β -error-rate is the long-run proportion of erroneous decisions to accept the point- H_1 as true/corroborated. Its complement, the $(1-\beta)$ -error-rate, is defined as the replication probability of a true effect (see notes 3 and 4).

² This isn't news, of course (Feynman, 1974; Meehl, 1978; Gigerenzer, 1998), but it has rightly been repeatedly identified as the probably most serious impediment to a cumulative science of human behavior (van Rooij & Baggio, 2020; Miłkowski, Hohol & Nowakowski, 2019; Irvine, 2021; Eronen & Romeijn, 2020; Eronen & Bringmann, 2021; Oberauer & Lewandowsky, 2019).

³ We write 'largely' because a replication study is subject to a regression effect, the size of which correlates inversely with the sample size. This means that smaller observed effect sizes are expectable in replication studies than in original studies (Fiedler & Prager, 2018). The basic factor here is the base-rate of replicable effects in a given field (Bird, 2020). Of course, the term 'field' is vague. Yet, the decisive problem is that quantifying the base-rate requires

entails that the original estimated observed effect size is typically based on heavily β -error-laden data. The chances of successfully replicating the original observed effect, therefore, are overly small to begin with.

Thus, a high degree of *empirically warranted confidence* cannot arise that a similarly generated effect size estimate matches the population effect size. For theory-construction purposes, however, this high degree of confidence is required, because an empirically adequate theoretical construct must model nothing other than the population effect. Lack of a high degree of confidence can thus be understood to arise from predominantly engaging with directional hypotheses. By converse reasoning, we reach the conclusion that theoretical progress in the behavioral sciences requires that empirical studies engage with a point- H_1 -hypothesis *by default*.

We start by explaining the RPS-specific integration of Frequentist and Bayesian statistical elements (Sect. 2), then argue that a composite hypothesis amounts to a barrier to theory-construction research (Sect. 3). We also address the meta-analytical integration of observed effects and explain why engaging with a point- H_1 -hypothesis is prone to require coordination between several labs (Sect. 4), then close with a brief summary of the main insights (Sect. 5).

2. Hypothesis testing in RPS, NHST, and BHT

2.1 The research program strategy

2.1.1 Data-based considerations

To a theoretician, empirical data (D) primarily provide relevant information to estimate the size of a population effect for which a theoretical construct shall be developed. When analyzing data statistically, researchers today rightly demand that the probability (p) of D in view of the H_0 must meet the criterion $p(D, H_0) < \alpha = 0.05$. Not only does this criterion indicate a statistically significant deviation of data from the H_0 . It also regularly provides a *necessary* condition for reviewers to recommend empirical findings submitted to a top journal for publication. (As a decision-criterion, the α -error-rate is comparable to the p -value; see our note 9.)

At the same time, behavioral scientists in their role as researchers or reviewers apparently find it acceptable that the estimated average β -error-rate of published studies reporting statistically significant findings is around 0.64 (Open Science Collaboration, 2015; Stanley, Carter, & Doucouliagos, 2018). If we assume a *medium* average observed effect size of $d=0.50$, itself a *very* generous assumption for the behavioral sciences (Linden & Hönemann, 2020), then the replication probability of *most* published findings would be a mere $(1-\beta)=0.36$ (Bakker, van Dijk & Wicherts, 2012). Seeing a replication crisis unfold in the

knowledge of the proportion of statistically *insignificant* findings given *all* findings in a field, including unpublished ones, a proportion that is highly uncertain.

behavioral sciences should therefore fail to surprise. What would surprise is if heavily underpowered findings did replicate. Yet, they don't.

A β -error-rate of 0.64 can seem unproblematic if a researcher's interest is exhausted by gaining empirical knowledge of a *directional* (non-random) effect. As the term says, a directional H_1 states the effect's positive or negative direction, but not its size. But as a directional H_1 avoids the effect's point-specification formally required to calculate the β -error-rate prior to data collection, the β -error-rate can only be calculated *post hoc*, after data are collected. A meaningful sense of 'sample size planning' therefore cannot arise. Thus, the amount of data that evidentially ground an empirical finding typically remains a *contingent* rather than a motivated quantity.

From a theory-construction perspective, however, this very quantity must be motivated. Indeed, an empirical study can yield readily *useful* information for theory-construction only if, prior to data collection, the α -error-rate, the sample size (N), and the theoretically predicted effect size are point-specified. These three parameters let NPTT infer the β -error rate. Respectively, pre-setting the α - and β -error-rates lets NPTT infer the minimum sample size (N_{MIN}) required to register an observed effect of *identical* size as the theoretically expected effect under these error-rates. (This is known as an *a priori* test-power analysis; Cohen, 1977). Since RPS endorses $\alpha=\beta<0.05$ as a *necessary* condition to obtain robust empirical findings, the amount of data that a theoretically useful empirical study must collect thus is a determined quantity.⁴ (We return to this in Sect. 2.2.)

Despite debate regarding the most adequate typology for replication studies (Machery, 2020), it is uncontroversial that a *direct replication* duplicates an original experimental setting, a *modified replication* alters it by incorporating alternative procedures or conditions, whereas a *conceptual replication* addresses the same theoretical idea using different techniques, techniques that change the experimental setting (APA, 2020). Hence, a direct replication is a data replication that must rely on the original study's N, the original α -error rate, the originally observed effect size, and a post-hoc calculated β -error-rate (Zwaan et al., 2018; Witte & Zenker, 2018). An original study's β -error-rate can thus be interpreted as the expected proportion of *failed* direct replications.

⁴ Cohen's (1977, 56) well-known convention ' $\beta=0.20$ conditional on $\alpha=0.05$, given no other considerations' is widely cited in the context of justifying the praxis of keeping data collection-cost manageable. Though Cohen had offered extremely weak reasons to support this convention, individual researchers nevertheless regularly endorse it as a *stopping rule* for data collection. At the collective level, however, this rule hinders theoretical progress, because the development of an empirically adequate theoretical construct presupposes that the observed effect size estimate is largely error-free ($\alpha=\beta<0.05$). Compared to the number of observations this requires, endorsing ' $\beta=0.20$ conditional on $\alpha=0.05$ ' as a stopping rule thus entails that the *planned* sample is invariably undersized ($N < N_{\text{MIN}}$).

Under this measurement-theoretical interpretation, the α - and β -error-rates express the distance from perfect measurement. Put differently, the α - and β -error-rates quantify the *measurement quality* an empirical setting achieves. According to the central limit theorem, increasing N continuously decreases both error-rates symmetrically. From a theory-construction perspective, therefore, what matters for an underpowered original study is a successful direct replication that increases the original study's sample size to N_{MIN} , or beyond. If the observed effect size in a well-powered direct replication then is at least *similar* to the observed effect size in the original study, this provides an empirical warrant to *increase* the degree of confidence in the assumption that the original study had made a discovery. In fact, this provides the only evidence-based way for a theoretician to increase confidence in the estimated population effect size.⁵

Having addressed the data/observations based on which the population effect size is *induced*, we proceed to address the hypotheses that state the estimated effect size either as *empirically observed* or *theoretically predicted*.

2.1.2 Hypothesis-based considerations

The H_0 -hypothesis typically states the mean of a random data-distribution ($m_0=0$) or the observed mean of the data-distribution in the control group ($m_1=x$). In behavioral science today, the third option, namely a point- H_1 that states a *theoretically predicted* mean, tends to be the exception. This should surprise because a theoretically predicted mean can be induced from a wide variety of available sources (Lakens et al., 2018). These sources include (i) minimum effect sizes observed in the past (e.g., in psychophysics); (ii) minimum important differences such as a perceived improvement on a depression scale (Button et al., 2015), or a minimum increase in monthly income sufficing to report “having more money,” or the effect size $d=0.50$ as a minimally important difference for a positive health outcome (Norman, Sloan & Wyrwich, 2005); (iii) disciplinary benchmarks such as the average observed effect of $r=0.21$ (i.e., $d=0.42$) for some 100 years of social psychological research (Richard et al., 2003); as well as (iv) meta-analytically estimated effect sizes, if corrected for the observed regression to the mean (Klein et al., 2018; Stanley et al., 2018).

⁵ This gain in confidence is subject to an important caveat. If an original study produces a *robust* finding, then, by definition, $\alpha=\beta<0.05$ holds. Even if similar error-rates hold in a replication study, however, an original observed effect size having been successfully replicated by independent efforts can only be a sufficient reason to state that *yet another*, namely a third data-replication is highly probable. Formally, this nevertheless fails to provide a sufficient reason to *increase* the evidentially warranted degree of confidence in the population effect size. For a *series* of studies with n findings, after all, the $(1-\beta)$ -error-rate is $(1-\beta)^n$. So, as n increases, the degree of confidence would decrease. In this context, we argue, the $(1-\beta)^n$ -measure is simply misleading (see Sect. 5).

The availability of sources sufficing to point-specify the theoretically predicted effect lets it appear questionable that behavioral science today predominantly engages with a directional H_1 . Indeed, similar sources would allow to predict a point- H_1 even *without* an explicit theory. In fact, should such sources prove insufficient to specify the H_1 as a point, one can always predict an *interval- H_1* , i.e., a two-point- H_1 . (An interval hypothesis is a restricted, more informative-rich directional hypothesis.) The interval's endpoints qualify as theoretical predictions, whereas the midpoint is a theoretical assumption. Given data, as distinct α - and β -error-rates apply at each endpoint, $\alpha=\beta<0.05$ should be associated to the interval's lower endpoint, that closest to the H_0 .

Generally, a theoretically predicted point-effect can be grounded either *inductively* using empirical observations, or *deductively* using theoretical considerations. In each case, gaining access to robust findings ($\alpha=\beta<0.05$) shifts the focus of inquiry from the *discovery context* to the *justification context* (Reichenbach, 1938; Witte & Zenker, 2017b). In the discovery context, researchers evaluate observations by relying on the α - and β -error-rates to calculate $p(D, H_0)$ and $p(D, H_1)$, that is, the probability of observed and yet more extreme data in view of two rivalling hypotheses.⁶ In the justification context, by contrast, what researchers evaluate are hypotheses. Here, $p(D, H_0)$ and $p(D, H_1)$, the values of which are established in the discovery context, serve to calculate the posterior probability of each hypothesis given data. This is conceptually equivalent to calculating the *relative support* that data lend to the H_0 and the H_1 . Although various controversial assumptions can underlie this calculation (see Sect. 2.3), there is no known alternative way of evaluating empirical hypotheses

⁶ The term ' $p(D, H_x)$ ' refers to the area under the bell curve, representing the density function for H_x (where $x=0,1$, with '0' for the control and '1' for the experimental group). As a natural language expression, this suggests 'the probability of D *in view of* H_x ', because as one calculates the probability of observed and yet more extreme data, one invariably *compares* the observed mean to the mean predicted by H_x . (In this, one does not conditionalize on the data, as ' $p(H_x|D)$ ' would suggest, nor does one calculate the joint probability of the hypothesis and the data, as ' $p(D \& H_x)$ ' would suggest.) Once these observations are transformed into a statistic (e.g., a t -value), $p(D, H_x)$ thus quantifies the probability of observing this statistic, or yet more extreme values of it. Given normally distributed data, moreover, the probability density is defined by the mean, m_x , and the standard deviation, s_x , the latter regularly *standardized away* (by setting $s=1$). Insofar as empirical measurement rightly associates with *non-zero* error-rates, then, each observed or more extreme data-point associates with a non-zero probability of representing an element of the data-distribution centered on $m_0=0$ or the data-distribution centered on $m_1=x$. So, to the extent that observations deviate from one of both means, the probability reduces that this mean adequately describes observations.

in evidence-based ways. Hence, a hypothesis evaluation in the *full* sense of ‘hypothesis testing’ is possible only in the justification context.

Conducting an RPS-hypothesis test is nothing other than to calculate the inverse probability that is known as the *likelihood ratio* (LR) (Edwards, Lindman & Savage, 1963):

$$LR_{10} = \frac{L(H_1|D)}{L(H_0|D)} = \frac{p(H_1) \times p(D, H_1)}{p(H_0) \times p(D, H_0)} \quad (1)$$

Whereas $p(D, H_0)$ and $p(D, H_1)$ are established inductively, from observations, selecting the prior probabilities, $p(H_0)$ and $p(H_1)$, is often considered a matter of subjective choice—and so allows for *asymmetrical* priors, $p(H_0) \neq p(H_1)$ —because all priors are eventually “washed out” by the evidence. (Given *limited* data collection resources, of course, the stress here is on ‘eventually.’) Yet, asymmetrical priors necessarily introduce a bias that make the hypothesis test unfair. A *fair* hypothesis test rather requires multiplying $p(D, H_0)$ and $p(D, H_1)$ with *symmetrical* point-specific prior probabilities, $p(H_0) = p(H_1)$. Under this condition, and given that the LR is a limiting case of the Bayes factor, $BF_{H1/H0} = p(D, H_1)/p(D, H_0)$, the LR becomes *identical* to the BF.

Following Wald (1947), RPS sets the corroboration-threshold for the H_1 to $LR_{H1/H0} = (1-\beta)/\alpha$, the ratio of test-power and statistical significance level. The corroboration-threshold for the H_0 is set accordingly to $LR_{H0/H1} = (1-\alpha)/\beta$. Given the demand for robust findings ($\alpha = \beta < 0.05$), the relevant instances of Wald’s thresholds, $LR_{H1/H0} > (1-\beta)/\alpha = 0.95/0.05 = 19$ and $LR_{H0/H1} > (1-\alpha)/\beta = 0.95/0.05 = 19$, thus yield necessary conditions to comparatively support, or corroborate, the first-mentioned point-hypothesis as opposed to the second. In brief, meeting Wald’s thresholds requires robust data. If this is the case, a *fair* hypothesis test turns into a *firm* hypothesis test.

The α - and β -error-rates, as we saw, quantify an empirical setting’s measurement quality, i.e., how error-prone the observed effect is, given data. RPS test-results therefore vary directly with the measurement quality. Since $\alpha + \beta = 2\alpha = 2\beta$, if N increases, then both error-rates decrease symmetrically, whereas Wald’s thresholds both increase. As desired, when increasing N , an RPS test-result asymptotically approaches the proportion of false positive and false negative errors (Krefeld-Schwalb, Witte & Zenker, 2018). Given a fair and firm hypothesis test, therefore, the value of the LR can be interpreted as the *strength of the empirical warrant* that the more supported point-hypothesis among considered point-hypotheses predicts the population effect size.

This is precisely the information a theoretician needs to identify in evidence-based ways the population effect size that a theoretical construct shall be developed for. For purposes of theory-construction, therefore, one would calculate the LR—to thus establish what RPS calls the *corroboration quality of hypotheses*—only if the expected proportion of erroneous decisions, or

interpretations, is $\alpha=\beta<0.05$. Empirical data meeting this condition achieve what RPS calls sufficient *induction quality of data*.

The motivation for introducing the terms ‘induction quality of data’ and ‘corroboration quality of hypotheses’, or the concepts thus denoted, is to (re-)connect applied statistical inference with general philosophy of science (Meehl, 1990). From a theory construction-perspective, after all, a useful statistical inference strategy must not only allow for an *induction* of a parameter-estimate from observations, but also for a *test* of a theoretically deduced prediction against new observations (Bandyopadhyay & Forster, 2011). Yet, most research efforts in behavioral science today are squarely inductive. Indeed, the conceptual link between induction and deduction appears to be severely underappreciated.

We proceed to detail the inference-elements that RPS adopts, or rejects, from the distinct yet related (Fienberg, 2006) statistical approaches that NHST and BHT are, and why the RPS-specific integration of these inference-elements yields an *all-things-considered* best statistical inference-strategy for theory-construction purposes.

2.2 Frequentist null-hypothesis significance testing (NHST)

2.2.1 Fisher’s NHST-version

Fisher’s (1959) NHST-version focuses on $p(D, H_0)$, the probability of data in view of the H_0 . Here, the point-specific (random) H_0 and the directional one-sided H_1 (or, in rare cases, an un-directed two-sided H_1) simply *describe* observations. By setting the mean observed effect in the control group to the mean of a random distribution, $m_0=0$, and comparing it to the mean observed effect in the experimental group, $m_1=x$, the expression $p(D, H_0)$ thus becomes useful to evaluate interventions that predict a *non-random* effect. An effective intervention, after all, implies $m_1 \neq 0$, while an ineffective intervention implies $m_1 \approx 0$. So, if $p(D, H_0) < \alpha = 0.05$, this warrants rejecting the H_0 as a faithful data-description; and if $p(D, H_0) > \alpha = 0.05$, this warrants *not* rejecting the H_0 . In brief, if $p(D, H_0) < \alpha = 0.05$, then observations can be interpreted as falsifying the H_0 (Popper, 1959), and as *potentially* supporting the H_1 .⁷

For instance, given a roughly normally distributed sample with an estimated standard deviation, s (aka a t -distribution), the decision to reject the H_0 , or not, rests on the probability that the proportion of measurement-scores that are assumed to be subject to *systematically uncontrolled* influences do not exceed the observed effect size. This probability is known as the t -value, $t = m_x / (s_x / \sqrt{N})$ (where

⁷ The term ‘hypothesis testing’ in ‘NHST’ easily misleads. In Fisher’s NHST-version, one first postulates the H_0 , then calculates $p(D, H_0)$, the probability that D *would* be observed if the H_0 *were* true. Yet, this licenses but a single inference, namely from ‘data have low probability in view of the H_0 ’ to ‘this indicates a H_0 -falsification’. By contrast, inferences serving to *test* the H_0 in the sense of calculating LR_{H_0/H_1} are unavailable, as is a corroboration measure for the H_0 or the H_1 (Gigerenzer, 2018).

$x=0,1$), and calculating that t -value is to conduct a t -test. Given $\alpha=0.05$ (one-sided), for example, and a sample of $N=21$ observations or data-points in each group ($N_{\text{TOTAL}}=N_1+N_2=42$), one finds $m_x > t=1.73$ (Faul et al., 2007). So, if the random H_0 describes observations faithfully, then fewer than 5% of all sampled measurement-scores exceed $t=1.73$.⁸ Thus, if $t=1.73$ holds for $N=21$, then Fisher's NHST-version licenses the inference from ' m_1 (as predicted by H_1) deviates statistically significantly from m_0 (as predicted by H_0)' to 'the evidence warrants rejecting the H_0 as a faithful data-description' at a 5% error-rate.

Yet, despite their statistical significance (read: despite being poorly described by a random distribution), a mere 21 measurement-scores would generate *sufficient* induction quality ($\alpha=\beta<0.05$), only if the observed effect were *large*, that is $d=(m_1-m_0)/s>1.00$. In behavioral science, however, the typical sample size today is an estimated $24<N<40$, and the typical observed effect size is an estimated $0.20\leq d\leq 0.50$ (Bakker, van Dijk & Wicherts, 2012). Given that RPS imposes $p(D,H_0)<\alpha=0.05$ and $p(D,H_1)<\beta=0.05$ as necessary conditions for a firm hypothesis test, *most* actual samples collected today thus simply are too small. This means that information required to estimate the population effect size *with a high degree of confidence* tends to be unavailable.

2.2.2 Neyman and Pearson's NHST-version (NPTT)

Making inferential use of the β -error-rate entails a transition from Fisher's NHST-version to its NPTT-version (Mayo, 1996; Mayo & Spanos, 2011). Given a pre-set d -value, $\alpha=\beta<0.05$ can in the first place suffice to define the necessary amount of data for an RPS-hypothesis test (see Table 1) because NPTT has exactly three degrees of freedom. Selecting values for any three of the following four parameters thus lets the final parameter-value "fall out."

(i) the α -error-rate: the probability of erroneously accepting that observed data deviate significantly from the H_0 (*false positives*)

(ii) the β -error-rate: the probability of erroneously accepting that observed data deviate significantly from the H_1 (*false negatives*); respectively the derived magnitude $((1-\beta)$ -error): the probability that observed data replicate, if the H_1 predicts the size of the population effect (aka *test-power*)

(iii) the *effect size* (e.g., d , r): the theoretically predicted, respectively the empirically observed effect, or correlation, between independent and dependent variables, when granting that the H_1 is true

⁸ Given $N=21$, a proportion of 5% seemingly implies 1.05 observations. When interpreted as a long-run chance, of course, the focal probability is meaningful without presupposing "observation fractions."

(iv) the *minimum sample size* (N_{MIN}), given values for the parameters (i) to (iii): the number of observations, or data-points, required to register a specific effect size under specific error-rates

Table 1: The minimum sample size (N_{MIN}) for a two sample t -test as a function of test-power ($1-\beta$) and effect size (d), given $\alpha=0.05$

$1-\beta$	d			
	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	88	34

To appreciate why NPTT necessarily brings a point- H_1 into play, consider that standardized effect size measures such as Cohen’s $d=(m_1-m_0)/s$, or the (related) point-biserial correlation measure $r_{bis}=d/\sqrt{d^2+4}$ —the two most frequently used effect size measures in behavioral science (Schäfer & Schwarz, 2019)—state the point-difference between the means of two data-distributions as standardized by the observed variance, s . In a randomized control trial, for instance, one data-distribution arises from how experimental participants behave “naturally,” making $d_{H0}=0$ the random value. The other distribution arises as the causal effect of a systematic intervention under background conditions, making $d_{H1}=x$ the alternative value. It follows that $d_{H1}=x$ can be interpreted as the smallest difference from random consistent with the assumption that a theoretical construct adequately predicts the intervention’s observed effect.

This interpretation conceptually presupposes the *theoretically predicted* mean difference, i.e., the difference between the means stated by the point- H_0 and the point- H_1 , m_1-m_0 , as distinct from the *empirically observed* mean difference. Crucially, a theoretical construct exclusively predicts the mean difference, m_1-m_0 , whereas the observed s expresses how accurately m_1-m_0 has been measured. And, for formal reasons, only if the theoretically predicted mean difference is point-specified prior to data collection can NPTT calculate the minimum sample size required for an observed mean difference of *identical size* as the theoretically predicted mean difference to count as a robust hypothesis-corroborating test-result. So, given the robustness of a test-result increases as the error-rates decrease, if the error-rates are pre-set, then N_{MIN} decreases as d increases, and vice versa.

In comparison, Fisher's NHST-version employs only the point- H_0 and the α -error-rate, though calling it ' p -value'.⁹ Here, the H_1 can be point-specified only *after* observations have been made. Unlike NPTT, therefore, Fisher's NHST-version cannot motivate a theoretically predicted effect *other than inductively*. Absence of a theoretical H_1 thus entails that neither the theoretically expected β -error-rate can be calculated, nor the point-effect that data corroborate, nor can a corroboration measure for the H_0 be had (Bayarri, Benjamin, Berger & Sellke, 2016).

The main consequence is that, in Fisher's NHST-version, one simply cannot *test* a theoretical H_1 against *new* (yet unobserved) data. This restricts statistical inferencing to the decision whether observed or yet more extreme data are sufficiently improbable in view of the H_0 . But this restriction is perfectly artificial. The α - and β -error-rates, after all, are nothing other than agreed-upon *conventions*, and so cannot pose a real problem. Moreover, we already saw that theoretical assumptions or empirical findings can suffice to point-specify the theoretically predicted effect. Indeed, as Lakens et al. (2018) observe, not point-specifying the H_1 has no excuse.

2.3 Bayesian hypothesis testing (BHT)

2.3.1 Stable test-results

In BHT, the term 'hypothesis corroboration' is interpreted as the relative support that a finite number of data-points lend to each element of a pair of hypotheses (Ly, Verhagen & Wagenmakers, 2016; Morey, Romeijn & Rouder, 2016). As the corroboration measure, BHT-proponents prefer the Bayes factor, $BF_{H_1/H_0} = p(D, H_1)/p(D, H_0)$, respectively $BF_{H_0/H_1} = p(D, H_0)/p(D, H_1)$. Generally, the larger LR_{H_1/H_0} or BF_{H_1/H_0} are, the stronger is the support that data lend to the H_1 , as opposed to the H_0 . (The opposite holds if H_1 and H_0 trade places in (1).)

⁹ Differences in origin, use, and interpretation generally keep from equating the α -error-rate with the p -value. In the context of an evidence-based *binary* choice ("accept" vs "reject"), however, all undeniable differences between both concepts lack *practical* bearing. The α -error-rate, originating in NPTT, states the long-run probability of committing a false positive error (aka type I-error). Originating in Fisher's NHST-version, by contrast, the p -value states the probability of observing actual or yet more extreme data, if the H_0 were true. Now, an evidence-based choice between the H_0 and the H_1 must consider the estimated probability of an *erroneous* choice. If the choice-criterion is the p -value, then this probability is based on data (*objective interpretation*); whereas if the choice-criterion is the α -error-rate, then this probability is based on a personal expected error-rate (*subjective interpretation*). Given an evidence-based choice is wanted in the first place, the personal expected α -error-rate must be set to a value *at least as large* as the p -value that provides the statistical significance criterion (e.g., $p=0.05$). In this way, then, the conceptual differences between the p -value and α -error-rate remain "hidden."

Following Jeffreys (1948), the threshold $BF_{H_1/H_0} > 10$ is conventionally interpreted as *strong* evidence for the H_1 . $BF_{H_1/H_0} > 30$ indicates *very strong* evidence, and $BF_{H_1/H_0} > 100$ indicates *extreme* evidence. In principle, of course, BHT-proponents could alternatively use LR_{H_1/H_0} and LR_{H_0/H_1} , as in RPS. A *fair* hypothesis test, after all, requires symmetrical prior probabilities of hypotheses, $p(H_1) = p(H_0)$. This implies that the BF is the LR (see equation 1, above).

By comparison, NHST estimates $p(D_x, H_i)$, rather than $L(H_i|D_x)$.¹⁰ This means that a corroboration measure is not *natively* available in NHST. So even data of arbitrarily high induction quality cannot allow NHST to corroborate H_i . Hence, NHST simply cannot evaluate the corroboration quality of hypotheses. It follows that NHST cannot “delete” empirically false assumptions or beliefs, nor can NHST “bury” dead theories (Ferguson & Heene, 2012; Hagger & Chatzisarantis, 2016). That said, RPS nevertheless relies on NPTT to estimate the induction quality of data before calculating the LR, thus *grounding* the corroboration quality of hypotheses in the induction quality of data.

BHT, by contrast, broadly *ignores* the induction quality of data. This creates a stability-problem that RPS avoids, but that BHT cannot resolve without becoming *equivalent* to RPS. To appreciate the problem, assume you have collected n data-points until a time point τ , and then calculate the BF (using any priors you like) to find that $BF_{H_1/H_0} > 30$, thereby finding the H_1 to be *strongly* supported (Jeffreys, 1948). For the same n data-points, however, also assume that the sample is so small, given the observed effect, that the β -error-rate *exceeds* 0.50 (see Table 1). It thus remains more probable than not that *new* data, yet to be collected after τ , support the H_0 rather than the H_1 .

Upon in fact collecting a sufficient amount of data of this kind, the initial test-result, $BF_{H_1/H_0} > 30$, will change, to instead favor the H_0 . So, the initial test-result was a *non-robust* finding. In this important sense, applying BHT *without* concern for N (read: without concern for the α - and β -error-rates) risks arriving at *unstable* hypothesis test-results.

The reason for incurring this risk is that, in BHT, data are interpreted as corroborating the numerator hypothesis in (1) as soon as the BF passes a corroboration threshold such as Jeffreys’s (1948). Yet, this interpretation ignores N , and the α - and β -error-rates demonstrably depend on N . The crucial weakness in BHT thus is that one cannot tell—except when going beyond BHT—whether a test-result *would* be stable if a sufficient amount of data of a certain kind *were* collected.

¹⁰ Notice the crucial difference with respect to the interpretation of ‘ D_x ’. If data are evaluated using the Fisher- or the NPTT-version of NHST, then ‘ D_x ’ denotes an *interval* under the normal curve, itself defined by actual as well as possible, yet more extreme observations. In RPS- and BHT-hypothesis testing, by contrast, ‘ D_x ’ denotes the *point-value* that represents the probability density of the mean of the H_0 -distribution, respectively of the H_1 -distribution.

A theoretician must find this weakness problematic, because any sample $N < N_{\text{MIN}}$ is individually insufficient for a *firm* LR-based hypothesis test. In fact, the LR and the BF themselves only provide a *quasi*-hypothesis test. The crucial additional step, which only RPS takes, is to consider the induction quality of data. BHT instead evaluates the support that data lend to hypotheses *already at the first observed data-point*, then adjusts the evaluation as per whichever data “come in.” This may suggest that BHT-proponents would “harden” the data by fiat—which is absurd (Lakatos, 1978). The one available alternative response entails a re-framing, such that BHT’s instability appears as a benefit. This explains why BHT-proponents emphasize BHT’s *immediate* responsiveness to data or its ability to learn from data *instantly*. (The stability problem, of course, remains the same.)

BHT’s stability problem is implicitly recognized in recent work on a *Bayes factor design analysis* (BFDA) (Schönbrodt & Wagenmakers, 2018). This analysis seeks to assist researchers in the planning-phase of studies that generate what the authors call “compelling” test-results. Like RPS, BFDA is sensitive to N , and thus specifically to test-power. Unlike RPS, however, BFDA fails to acknowledge the conceptual link between induction and corroboration, the very link that motivates why Wald’s corroboration thresholds are sensitive to N . That this link is unacknowledged in BFDA shows when, in lieu of an *objective*, data-based criterion for the stability of test-results, Schönbrodt & Wagenmakers (2018, Table 1) again resort to Jeffreys’s (1948) BF-thresholds. Jeffrey, however, did *not* ground his BF-thresholds in the induction quality of data; he rather selected them perfectly *ad lib*. Jeffrey’s BF-thresholds thus *ignore* the measurement quality of an empirical setting, hence making them subjective thresholds.

In an RPS-hypothesis test, by contrast, what motivates the $\alpha = \beta < 0.05$ -constraint on induction quality of data, as well as the shape of Wald’s thresholds ($LR_{H1/H0} > [(1-\beta)/\alpha] = 19$; $LR_{H0/H1} > [(1-\alpha)/\beta] = 19$), is the *need to control the proportion of erroneous data-based acceptance- or rejection-decisions* regarding an empirical hypothesis. Proposed during WWII as a control mechanism to improve the quality of ammunition production-output, Wald’s thresholds were originally applied in *sequential testing*, where a minimum sample size is not determined beforehand. As is the case in BHT, rather, data are continuously evaluated until sampling stops. When Wald’s thresholds are instead used *non-sequentially*, they acquire the desirable property of tracing a test-result’s error-rates, and thus trace an empirical setting’s measurement quality. Since this property entails that one can control the proportion of erroneous decisions asymptotically for the H_0 and the H_1 , Wald’s thresholds are data-based, objective thresholds.

2.3.2 Fair and firm hypothesis tests

Wagenmakers and colleagues BHT-version is a sophisticated extension of Jeffreys’s (1931; 1948) approach. Not only did Jeffrey “distinguish [sic] sharply between problems of parameter estimation and problems of hypothesis testing” (Ly, Verhagen & Wagenmakers, 2016, 20). He also understood that Fisher’s

(1925) p -value can at most serve to evaluate an empirical, rather than a theoretical parameter. A Fisher-style test, after all, offers but a single inductive inference according to which the p -value of observed data falsifies the H_0 . As we saw, however, this inference is insufficient to *test* the H_1 .

Conceptually, an *induced* H_1 never amounts to more than a hypothesis that is best-fitted to data. An H_1 *to-be-tested*, by contrast, must do more than merely describe “old,” already observed data. It must also predict *new* data. A reasonable criterion for a hypothesis test is that, under initial and boundary conditions, the point-specific theoretical H_1 is logically deduced from a theory, T (Hempel & Oppenheim, 1948; Meehl, 1990, 1967; Robinaugh et al., 2020).

This criterion presupposes a hard conceptual distinction between *induction* and *deduction*, because if the theoretically deduced H_{DED} were *non-distinct* from the empirically induced H_{IND} , then data already used once, to induce the H_{IND} , could be re-used, to probabilistically confirm the H_{DED} . Such “data-recycling” is what BHT fully accepts, indeed embraces. Yet, this stance simply *levels* the induction vs. deduction-distinction. In BHT, therefore, a test of a H_{DED} that exclusively predicts new data, as in RPS, is *conceptually* impossible.

Usage of the BF in BHT thus differs importantly from how the LR is used in RPS. Committing to induction vs. deduction as a hard distinction entails that an empirical hypothesis must be predicted *not* as an inductive consequence of D , but as deductive consequence of T . And, this deductive consequence must be tested against new data that allow for robust test-results ($\alpha=\beta<0.05$), with the right corroboration measure being not $p(D, H_0)$ to falsify H_0 , but LR_{H_1/H_0} to probabilistically (dis-)confirm H_1 relative to H_0 .

These constraints suffice for a *firm* hypothesis test. But a *fair* hypothesis test requires yet more. As per how equation (1) relates $p(H_x)$ to $p(D, H_x)$, after all, the expression ‘ $BF_{H_1/H_0}=p(D, H_1)/p(D, H_0)=x$ ’ can be interpreted as: ‘in view of data, the numerator-hypothesis is x times as probable as the denominator-hypothesis’. (This agrees with Jeffreys’s intended interpretation.) Independently of having collected data, however, this interpretation makes *asymmetrical* priors, $p(H_1) \neq p(H_0)$, equivalent to a preference or *bias* pro/con the H_1 - or the H_0 -hypothesis that necessarily taints the test-result’s strength. In brief, asymmetrical priors make a hypothesis test *unfair* (Witte & Zenker, 2017b; Krefeld-Schwalb, Witte & Zenker, 2018).

In sum, a fair and firm hypothesis test (i) point-specifies the H_0 (typically as $d_{H_1}=0$) and the H_1 ($d_{H_1}=x$); (ii) uses *symmetrical* priors to avoid biasing the test-result; (iii) pitches the H_0 and the H_1 against *new* data; and (iv) uses samples that meet N_{MIN} (given $\alpha=\beta<0.05$, $d_{H_1}=x$, $d_{H_0}=0$). Each of these four constraints is motivated directly from the *theoretical* need for fallible knowledge of the population effect size for which a theoretical construct is to be developed. Outside of these constraints, the statistical inference strategy that BHT is cannot *natively* corroborate point-hypotheses fairly and firmly. Whereas inside these constraints, BHT *is* RPS.

3. Composite hypotheses as barriers to theory-construction

3.1 Pure induction and the true population effect

Based on the statistical regularity known as the *central limit theorem*, an often-voiced reason against point-specifying the H_1 is that, if N increases, then a maximum likelihood-estimation is guaranteed to induce—in the long run and with ever increasing probability—the point-value upon which the empirically observed effect and the population effect converge. Computer simulations, which invariably rely on the central theorem, readily demonstrate such convergence. Were this the full story, then experimental or observational studies could simply *induce* the population effect from data, making a theoretically motivated point-effect size superfluous.

Before a simulation can be run, however, the population effect must be *defined*, rather than being observed. The convergence can therefore be proved only in this trivial sense. Given actual rather than simulated conditions, moreover, the *validity* of an inductive parameter is limited to the specific samples, tests, experiments, dependent variables, time-points, etc., that provide the evidential basis for the parameter (Borsboom, Mellenbergh, & van Heerden, 2004). In terms of the validity of a test-result, therefore, a parameter that is induced from old data is *conceptually non-equivalent* to a theoretical parameter that is tested against new data.

Under actual conditions, observations also tend to fluctuate. This already holds if only random error-fluctuations are at play, which must always be assumed. Worse yet, Linden & Hönekopp's (2021) arguably *devastating* meta-meta-analytical results show for behavioral science that, if the observed effect size is large enough to be predicted by a theoretical construct, then one typically cannot induce a *sufficiently homogenous* parameter, with small variance (see Olsson-Collentine, Wicherts, & van Assen, 2020; Schauer & Hedges, 2020). Conversely, sufficiently homogeneous induced parameters typically remain too small for a theoretical construct to predict them, because such small effects are overlain by the measurement-error. Yet, a *firm* hypothesis test requires a homogenous induced parameter, because the heterogeneous variant entails different causal influences upon human behavior to be at work simultaneously, influences that must be explained by different theoretical constructs (Bryan, Tipton & Yeager, 2021).

At the initial stage of a research program, therefore, where theoretical knowledge is by definition insufficient and samples tend to be (too) small, a hypothesis test normally fails to be a *meaningful* act of inference. Yet, in behavioral science, this insight appears to lack uptake today. For a test-result to be publishable, it broadly suffices that statistically insignificant data are *avoided*. The “optimum” tool to this end arguably is a (one-sided) H_1 that predicts an unspecified directional difference from the H_0 . In observational and experimental studies alike, however, assuming the presence of random error-fluctuations implies that some statistically significant effects are *theoretically irrelevant*. Add to this that a directional H_1 (where the β -error-rate is undefined) can never yield a

stronger test-result than a *preliminary discovery*, defined as $p(H_0, D) < \alpha$ -error-rate. So, a preliminary discovery fails to provide an adequate basis for erecting a theoretical construct in evidence-based ways.

In behavioral science, nevertheless, a preliminary discovery *at object-level* is regularly treated as if its relevance for theory construction did not require separate argument. Yet, to assess the relevance of a preliminary discovery rightly remains the genuine purpose of *meta-analytical* research. Here, preferably homogenous, topically related object-level test-results (which increasingly are preregistered today) are pooled to derive a weighted point-average effect, d_{HI} . In RPS, one would next seek to corroborate d_{HI} using the LR and Wald's thresholds. For a corroborated d_{HI} to become part of a research program additionally requires a *theory-core* (Lakatos, 1978; Meehl, 1990). Yet, since a theory-core is typically unavailable at a research program's initial stage, the search for law-like regularities (that a theory-core subsumes) can be inferentially meaningful only at a research program's subsequent stages (Witte & Zenker, 2016; 2017a; 2017b; Krefeld-Schwalb, Witte & Zenker, 2018).

These considerations cast strong doubt upon the claim that the *telos* of in behavioral science research today is in any important sense theoretical. Rather, when going by the discussion section of a typical journal publication, authors regularly (and falsely) equate 'making a preliminary discovery' with 'conducting a hypothesis test' (Greenland et al., 2016), and generally seem to overestimate the relevance of a preliminary discovery for theory construction. But unless a preliminary discovery is replicated several times using robust data, a theoretician seeking to develop an empirically adequate construct should simply *ignore* it.

3.2 Selective publishing and Popper's falsification principle

To gain the undistorted view upon behavioral science research that the praxis of selectively publishing statistically significant test-results makes impossible today (Rothstein, Sutton & Borenstein, 2005), a publication decision should arguably be *independent* of whether a test-result is statistically significant, or not. Presently, however, *publication bias* implies that statistically insignificant test-results are broadly excluded from publication (aka file drawer problem; Rosenthal, 1979), and so cannot easily inform meta-analytical research efforts either. Unless one adequately corrects for this bias, which proves to be very difficult, the meta-analytically estimated population effect is prone to be overestimated. This contributes to explaining why published findings in behavioral science typically fail to replicate.

The perhaps most sophisticated counter measure against selective publishing is *results blind manuscript evaluation* (RBME) (Locascio, 2017; 2019). (For other measures, see Berlin & Ghersi, 2005; Chambers, 2013). RBME bifurcates the publication decision process into a results-blind review stage, followed by a results-transparent stage. Obviously, if the decision from the results-blind stage is maintained, editors' and reviewers' preferences towards statistically significant test-results can surface only in the results-open stage,

where they matter less or fail to matter. So, RBME is conducive to reducing the negative impact of publication bias upon meta-analytical research.

In Fisher's NHST-version, however, the statistical-inferential theory that behavioral scientists predominantly employ today, RBME interestingly *fails* to support the development of empirically adequate theoretical constructs. For this purpose, indeed, RBME is effective only in RPS. To explain, Fisher's NHST-version rests on Popper's (1959) *falsification principle*, stating that a hypothesis can be falsified, but not verified. (Slogan: 'in Fisher's NHST-version, one can never accept the H_0 ; one merely retains or fails to reject it'). The falsification principle thus implies that a statistically significant test-result is informative with respect to falsifying the H_0 , whereas a statistically insignificant test-result is to this end perfectly uninformative.

In NHST, falsification thus is *asymmetrical*: one can falsify the H_0 , but not the H_1 . In RPS, by contrast, falsification is *symmetrical*, because RPS assesses the induction quality of data regardless of whether a test-result is statistically significant. Therefore, as long as a test-result is robust, RPS allows to falsify or corroborate the H_0 *and* the H_1 . Thus, particularly a statistically *insignificant* test-result can corroborate the H_0 in RPS, but not in NHST.

In the context of NHST, moreover, the praxis of selective publishing cannot be properly understood as an exclusive consequence of career-related *external* publication incentives, because Popper's asymmetrical falsification principle implies that the praxis already arises *internally*. Insofar as the falsification principle explains why a statistically significant non-random test-result amounts to a H_0 -falsification, the falsification principle explains why such test-results are potentially relevant for theory construction. Yet, the falsification principle also implies that a statistically *insignificant* test-result *should* remain unpublished, because the publication of uninformative test-results amounts to an epistemically pointless act for a falsificationist.

Given that the asymmetric publication praxis in the behavioral sciences today reflects Popper's falsification principle, selective publishing can thus be explained to result from NHST being based on Popper's deterministic logic of falsification. Within Popper's own logic, however, it is not readily meaningful to understand the non-publication of H_0 -corroborating test-results as a result of publication *bias*. The term 'bias' becomes meaningful only if Popper's deterministic logic of falsification is replaced with a probabilistic logic of verification, a logic offering a formal measure to falsify or corroborate the H_0 and the H_1 *symmetrically*, as in RPS. Within Popper's logic, the praxis of selective publishing rather is evidence that behavioral science heeds the falsification principle. A theoretician must therefore ultimately reject this principle.

3.3 Induction

Probabilities are generally useful to estimate a parameter from observations. Yet, to develop a theoretical construct, what counts are *likelihoods*. Provided $N \geq N_{\text{MIN}}$, after all, only a likelihood-ratio test can corroborate a point-specific H_0 - and the

H_1 -hypothesis fairly and firmly. As we saw, alternative evidence-based ways of warranting a *high* degree of confidence in a population effect size estimate are unknown. And, anything less than a high degree makes it unreasonable to even begin developing a theoretical construct. For this reason, then, hypothesis testing in RPS requires the point- H_1 that Fisher's NHST-version cannot provide, and that BHT normally will not provide.

As we also saw, not only is it possible to point-specify the H_1 . It is also impossible to overstate its theoretical significance: "[...] not [point-]specifying a SEOI [smallest effect of interest] for research questions at all will severely hinder theoretical progress" (Lakens et al., 2018, 267). Yet, to most behavioral scientists engaged in statistical inference today—proponents of Wald's (1947) sequential testing approach and Royall's (1997) likelihood paradigm excepted—this statement presumably echoes a fringe position. In fact, many prominent researchers have declared a point- H_1 to be *infeasible*. Mayo and Spanos (2011, 175), for instance, mention the "highly artificial case of two point hypotheses." Similarly, Rouder et al. (2009, 229) write: "[w]e assumed that the alternative [hypothesis] was at a single point," an assumption that is "too restrictive to be practical."

The alleged "problem with [likelihood testing] is that the [point-]exact alternative effect size δ [i.e., $d_{H1}=x$] is never known beforehand," wherefore "in Bayesian statistics, this uncertainty about δ is addressed by assigning it a prior distribution" (Verhagen & Wagenmakers, 2014, 1461). Such claims, however, should give pause, because the point-specification of the random H_0 is widely accepted as unproblematic today. For good reasons, too, because $d_{H0}=0$ quantifies a central *theoretical* assumption regarding theoretically *non*-modelled compensating influences (to the left and right of the H_0) thought to lie beyond the causal reach of an intervention. Yet, the random H_0 is point-specified in *exactly the same way* as the theoretical H_1 . And, again, without a point-specific H_1 , one cannot evaluate observations in view of a theoretical hypothesis.

Thus, rather than getting to test a theoretical construct against new data, one must rest content, as BHT does, with *inducing* a parameter-estimate from data. For theory-construction purposes, however, already Fisher's NHST-version shows that an induced parameter-estimate is of very limited theoretical use-value. Given a point- H_0 , after all, even if data prove less probable than the α -error rate, as long as the β -error rate is undefined, one can at most obtain a *preliminary* H_0 -falsification. This provides a warrant to *reject* the H_0 . But since one here treats merely the data-distribution centered on the H_0 , while ignoring the distribution centered on H_1 , a warrant to *accept* the H_1 remains absent.

Moreover, the properties of the H_0 -distribution depend on N , because N determines the confidence interval (CI), centered on the H_0 , indicating that data deviate significantly from the H_0 . As a direct function of N , therefore, *different* CIs are associated with the H_0 . Rather than motivating these CIs substantially, however, they derive from the statistical theory underlying a Fisher-test. But a Fisher-test measures only $p(D, H_0)$, and so treats only the H_0 . In brief, all this is

induction *pure*. A theoretician seeking to develop an empirically adequate theoretical construct, and to test it fairly and firmly, must want more.

3.4 The priors

Empirical data generally suffice to reduce a researcher's uncertainty regarding a true hypothesis *maximally*, only if the entropy of the empirical setting is maximal, that is, if the setting is "maximally noncommittal with regard to missing information" (Jaynes, 1957, 171). A hypothesis test can satisfy this constraint only if the prior probability of the H_0 and the H_1 is set to $p(H_0)=p(H_1)=0.50$, as in RPS (Attnave, 1959), because asymmetrical priors, $p(H_1) \neq p(H_0)$, necessarily bias the test. Particularly, if that bias is *empirically false*, then an asymmetrical distribution is also guaranteed to waste resources, because each bias-consistent data-point must first be corrected by a bias-compensating data-point, until a symmetrical prior distribution obtains. Only then can new data bear *properly* on each hypothesis, so that the increase in empirical knowledge per amount of invested resources is maximal.

This provides a sufficient reason to *eliminate* the prior distribution of possible H_1 -hypotheses that is characteristic of BHT, making a motivated choice among different prior distributions superfluous. This counts, because quantitative differences with respect to the prior distribution of possible H_1 -hypotheses can lead to qualitative differences with respect to corroborating the H_1 . In extreme cases, indeed, changing a hypothetical parameter's prior distribution (e.g., from a Cauchy to a normal distribution) may even let the *same* data once support the H_0 and once the H_1 , entirely because of the (subjective) probability weight assigned to the hypothetical parameter prior to data collection (see Bem, Utts & Johnson, 2011; Wagenmakers et al., 2012).

Yet, the only reason why the corroboration-verdict can change in the first place is that, in BHT, "the uncertainty about [the alternative hypothesis] δ is addressed by assigning it a prior distribution" (Verhagen & Wagenmakers, 2014, 1461), rather than a point. To justify this assignment, BHT-proponents regularly allege that a point- H_1 is an overly strong assumption that unduly reflects researchers' *subjective* beliefs. This justification, however, becomes self-defeating when fully appreciating the subjectivity intrinsic to BHT's prior probability-ratio, $p(H_1)/p(H_0)$, a ratio that is deduced from the theoretical parameter's distribution prior to data collection. In equation (1), after all, the posterior probability-ratio is the product of $p(H_1)/p(H_0)$ and the ratio of the probabilities of data in view of two rival hypotheses, $p(D, H_1)/p(D, H_0)$. Whether one motivates $p(H_1)$ and $p(H_0)$ by assumption, by observation, or by speculation, BHT invariably specifies $p(H_1)$ and $p(H_0)$ *distributively*. Both distributions nevertheless describe a fixed "point-ratio" that is indispensable to calculate the posterior probability-ratio. So, if $p(H_1)/p(H_0) \neq 1$, then *in virtue of the probability weight*, the point-ratio necessarily influences whether this or that hypothesis is corroborated by data.

Of course, BHT's point-ratio easily seems more "forgiving" than the ratio of the point-specific H_0 - and H_1 -hypotheses in RPS. Yet, both ratios are of the

same formal strength. In fact, BHT lends itself of point-specification as much as RPS does. Alas, while BHT point-specifies possible values of $p(H_x)$ as relative probabilities that follow specific distributions, RPS point-specifies the H_0 and the H_1 themselves. The crucial difference thus is the *type of information* specified. Whereas BHT specifies the comparatively “softer” $p(H_x)$, RPS point-specifies H_x itself, either by deriving H_x from a theory or by inducing H_x from robust antecedent test-results. Insofar as what BHT point-specifies thus are relative probabilities, however, a distributive specification simply leaves the “theory-part” of (1) *vague*, the very part that RPS specifies as a point-precise theoretical hypothesis. This explains why RPS can eliminate the prior probability distributions that BHT requires.

Given that empirical measurement is subject to error, moreover, what BHT’s probability distribution for $p(H_x)$ represents is the *measurement error*. Since hypothesis testing must rely on measurement, of course, both BHT and RPS incur these error-distributions necessarily. What the measurement error reflects, however, are researchers’ contingent measuring experiences. At any rate, the measurement error cannot be motivated from formal or practical difficulties with formulating point-hypotheses. So, although empirical measurement must rely on probability distributions, it is false that hypothesis-specification must do the same.

4. Default point-hypotheses and resource pooling

4.1 Research programs, default H_1 -hypotheses, and meta-analysis

Although Fisher’s NHST-version merely provides an unspecified, directional H_1 that renders a priori test-power unknown, a Fisher-test (i.e., a test against the H_0) remains the dominant statistical tool in behavioral science today to warrant the publication of empirical findings. But not only does a Fisher test thus encourage researchers to avoid a point- H_1 . It also encourages them to collect (what test-power considerations identify as) undersized samples, and to work in near-isolation from labs doing related research. The dominance of Fisher’s NHST-version thus contributes to explaining why published test-results are typically underpowered and why the evidential basis for developing an empirically adequate theoretical construct is therefore broadly absent.

Arguably, if the behavioral sciences were aligned to the theoretical goal of gaining empirically warranted, fallible knowledge of population effects, then something like RPS should already be normal today. Currently, however, research programs in the behavioral sciences are underdeveloped, or simply do not exist. Of course, any research program grounds partially in speculation. Indeed, a laborious process of exploratory data collection is required before an NPTT-style meta-analysis can yield a parameter-estimate of the population effect size that would serve as a *default* point- H_1 -hypothesis (Anderson & Maxwell, 2016).

Conceptually, this default point- H_1 expresses a researcher’s initial ignorance of the population effect size. Only once a formal theory-core is developed can the identity conditions for a theoretical construct be settled (Lakatos, 1978). This in place, initial and boundary conditions jointly imply

competing point-predictions (H_0 , H_1) that can be statistically falsified or corroborated. In fact, even given a mere *qualitative* theory (which by definition lacks a theory-core) the theoretically predicted effect size can nevertheless be specified as a percentage (e.g., 50/100 cases), and be tested against, say, a medium effect (65/100 cases).

Besides motivating the theoretically predicted effect from available empirical sources such as listed in Sect. 2.1, general inductive considerations suggest that the *median* effect size of published studies in behavioral science, $d=0.50$ (aka a medium effect; Cohen, 1977), may provide a reasonable default H_1 -hypothesis (Schäfer & Schwarz, 2019). Yet, if the true population effect would be $d=0.50$, then the typical sample size of experimental studies in the behavioral science, $N_1=N_2=30$, implies that a (one-sided) t -test under $\alpha=0.05$ registers $d=0.50$ at test-power of $(1-\beta)=0.61$. So, in the long run, 39 in 100 such tests would expectably fail to replicate $d=0.50$, because similar test-results are insufficiently powered.

Further, a top-tier journal typically accepts empirical findings about an identical effect for publication, only if a submitted manuscript reports at least three independent experimental studies with statistically significant test-results. Given that the probability of obtaining a statistically significant test-result in n independent studies is $(1-\beta)^n$ (Francis, 2012), if $(1-\beta)=0.61$ holds for a typical individual study, then the *joint* replication probability of three typical studies with statistically significant test-results is $0.61^3=0.23$. *Ceteris paribus*, if the samples of three individually underpowered studies were instead pooled into a *single* study, test-results would be far more robust. Given $\alpha=0.05$ (one-sided) and $(1-\beta)=0.61$, for instance, detecting $d=0.50$ requires $N_{\text{MIN}}=N_1=N_2=47$. Whereas three times that sample, $N=141$, suffices to detect $d=0.50$ under $\alpha=0.05$ and $(1-\beta)=0.99$, respectively $\alpha=0.02$ and $(1-\beta)=0.98$. (Recall that, by increasing N , the error-rates decrease symmetrically).

A theoretician would find similarly small error-rates attractive. With respect to theory construction, however, that a statistically significant well-powered original observed effect has been replicated seemingly *fails* to warrant an increased degree of confidence that effects observed in the original and the replication study *converge* upon the population effect. After all, the degree of confidence in a series of n independent studies with statistically significant test-results would likewise decrease with n .

For instance, even if each individual study reports the same observed effect size at test-power of $(1-\beta)=0.99$, $(1-\beta)^n$ implies that already only six such studies generate *insufficient* joint test-power of $(1-\beta)=0.94$. By parity of reasoning, this result also pertains to the estimated (non-random) effect size that a *meta-analysis* derives from a large sample of independent object-level studies. With each additional study being included into the meta-analysis, therefore, an *ever-decreasing* degree of confidence would be warranted that the meta-analytically estimated effect size matches the population effect size.

Given that selective publishing is the norm, the sample of object-level studies grounding a meta-analytical population effect size estimate must itself be thought of as originating from a population of studies that predominantly report statistically significant test-results. And, the same holds for the pattern of three statistically significant test-result that top-tier journals require for publication. But if underpowered object-level studies are the norm, then it is statistically unavoidable to obtain at least some *insignificant* test-results. It is therefore a priori more likely than not that the pattern required for publication originated in a population of studies from which statistically insignificant test-results were *excluded*.

Several recent *meta-meta*-analyses of behavioral science findings show that the homogenous partition of observed effects typically comprises *small* effects ($0.01 \leq d \leq 0.20$), whereas relatively infrequently observed *medium* to *large* effects ($0.50 \leq d \leq 1.00$) are typically highly heterogeneous (Olsson-Collentine, Wicherts, & van Assen, 2020; Schauer & Hedges, 2020; Linden & Hönemann, 2020). We already saw that replication studies that observe a large(r) heterogeneous effect are *theoretically ambiguous*, because such effects require a differentiated explanation of observed heterogeneity (which is hard to come by); whereas replication studies that observe a small homogeneous effect are *theoretically uninteresting*, because the observed effect is overlain by the measurement-error.

The predominance of small object-level observed effects thus entails that the evidential basis for a meta-analytical effect size estimate necessarily includes a sizable share of test-results the statistical significance of which is owed exclusively to the α -error-rate. Thus, at object-level, false positive test-results are erroneously identified as true positive test-results. It follows that the object-level studies from which a meta-analysis samples already *overestimate* the observed effect size, the overestimate itself being heterogeneous.

For this reason, the meta-analytical population effect size estimate requires a correction factor that is sensitive to the number of *unpublished*, statistically *insignificant* test-results that the praxis of selectively publishing statistically significant test-results typically excludes from publication (Maxwell, 2004). Although state-of-the-art meta-analyses do seek to correct for excluded test-results (normally via an e-mail request that a researcher check her file-drawer), the full extent of selective publishing is simply unknown. So, the number of insignificant results that a sound meta-analysis *must* include is equally unknown.

The undeniable formal soundness of these considerations not only suggests that current meta-analytical effect size estimates provide severe misestimates. These consideration also suggests that, for purposes of theory construction, $(1-\beta)^n$ is a particularly poor formal measure, because it supplies a highly *misleading* assessment of how trustworthy an inductive parameter estimate is in view of data. Particularly because the selective publication of underpowered test-results is the norm today, an adequate formal measure for this purpose must

allow researchers to *pool* underpowered test-results into a *global* point-hypothesis test. We now turn to how RPS enables this.

4.2 Underpowered test-results and log-LR addition

As we saw, not only does NPTT allow researchers to plan a well-powered study by calculating N_{MIN} . The required point-specification of the H_1 can also present a hardly more challenging task than a point-specification of the H_0 . Since the rarity of a point- H_1 in behavioral science research today explains why test-results are typically underpowered, this suggest that the behavioral sciences as a whole are primarily aligned to the discovery of statistically significant effects using a minimum amount of resources (Gervais et al., 2015).

In psychology, for instance, by far the largest behavioral science field, N_{MIN} regularly exceeds the typical sample size, $24 < N < 40$, *massively* (Bakker, van Dijk & Wicherts, 2012). In fact, given the estimated median observed test-power in psychology, $(1-\beta)=0.35$ (Bakker et al., 2012), if we assume $\alpha=0.05$ and a generous $d=0.50$, then we find $N_{\text{MIN}}=N_1=N_2=14$. A *firm* evidence-based decision regarding $d=0.50$, however, requires $\alpha=\beta<0.05$, thus implying $N_{\text{MIN}}=N_1=N_2=88$. A typical behavioral science study would thus appear to lack about 84% of N_{MIN} .

NPTT's α - and β -error-rates, of course, are meaningful only for observed data, as well as for possible, yet more extreme data. Moreover, the scope of NPTT's intended application includes neither the integration of individual test-results, nor the corroboration of hypotheses (Bayarri et al., 2016, 100f). Rather, NPTT's intended application is to estimate the mean effect size that best *describes* observations. To this end, observations are associated with conventionally agreed-upon α - and β -error-rates. And, if only relatively few confidence intervals (CIs) include a zero test-result, then the estimated observed mean can be interpreted as a statistically significant effect (aka "discovery"), and a non-significant effect otherwise.

To construct, or corroborate, a theoretical construct in evidence-based ways, however, NPTT-results are prone to mislead, because the α - and β -error-rates ground a firm decision as to the best data-describing mean only for an *individual* study. Yet, for a series of studies, we saw that the *global* $(1-\beta)$ -error-rate decreases as the product of the $(1-\beta)$ -error-rate of each individual study in this series. A global test-result for a series of statistically significant test-results would therefore achieve *sufficient* test-power ($(1-\beta)>0.95$), only if the $(1-\beta)$ -error-rate for each individual study were already extremely close to 1. Otherwise, the global $(1-\beta)$ error-rate quickly drops below 0.50. This holds for any value of N .

As a statistical magnitude, N does provide a crucial magnitude to distinguish a data description from a hypothesis test, because for a data description the error-rates are set conventionally, whereas for a hypothesis-test the error-rates vary with N . That the global $(1-\beta)$ -error-rate decreases as N increases nevertheless renders the global $(1-\beta)$ -error-rate an *inadequate* formal measure to assess the evidential basis for theory construction. If we assume a series of n

statistically significant, well-powered studies each of which arrives at the *same* estimated observed effect size, $d=x$, then the behavior of the global $(1-\beta)$ -error-rate *misleadingly* entails that the empirically warranted confidence in $d=x$ as a population effect size estimate decreases as n increases. Yet, the *opposite* conclusion holds. For a global point-hypothesis test, then, the global $(1-\beta)$ -error-rate is an inadequate formal measure.

With undersized samples the norm in behavioral science, an adequate formal measure for global point-hypothesis test must particularly allow researchers to *pool* test-results from a series of individually underpowered studies. To this end, RPS employs the *log-likelihood-aggregate* (log-LR), that is, the log-addition of $L(H_1|D)/L(H_0|D)$ for each study in this series. Computer simulations show that calculating the log-LR for a series of individually underpowered studies *compensates* for each individual test-result failing to be robust (Krefeld-Schwalb, Witte & Zenker, 2018). So, provided the global log-LR-sum passes the relevant instance of Wald's thresholds, even underpowered test-results can support hypotheses, and ever more firmly as N increases.

In praxis, therefore, although a single lab may regularly lack sufficient resources to collect N_{MIN} , several labs can in principle collect N_{MIN} jointly. Where undersized samples are normal, a global test-result thus requires collective efforts (Open Science Collaboration, 2015; Etz & Vandekerckhove, 2016). *Immediate* trust in the global test-result nevertheless requires altogether abandoning the selective publication of statistically significant test-results. So, until this praxis has ceased, e-mail requests to check one's file drawer for unpublished non-significant test-results are an evidentially reasonable measure, because the global log-LR should include such test-results.

In sum, insofar as behavioral scientists can, and regularly do point-specify the H_0 , as $d_{H_0}=0$, they can likewise point-specify the theoretically predicted effect, as $d_{H_1}=x$. A sufficient reason to motivate particularly $d_{H_1}=0.50$ as a default theoretical H_1 -hypothesis is it allows calculating test-power prior to conducting a hypothesis test. Of course, if the observed effect size exceeds the theoretically predicted effect size (wherefore the a priori value of N_{MIN} exceeds the post-hoc value), labs are guaranteed to "waste" resources. Any associated sunk cost, however, are perfectly acceptable, because the only available alternative strategy is equivalent to resource-thrift dominating a low $(1-\beta)$ -error-rate. From a theory-construction perspective, alas, this preference order is detrimental (Bayarri et al., 2016, 100).

5. Conclusion

Based on Wald's (1947) sequential testing approach, RPS integrates select elements from Frequentist and Bayesian statistical inference. RPS's main goal is to support the development of empirically adequate theoretical constructs in evidence-based ways. An RPS-hypothesis tests therefore relies on Wald's thresholds, $LR_{H_1/H_0} > (1-\beta)/\alpha$ and $LR_{H_0/H_1} > (1-\alpha)/\beta$, thresholds that trace the induction quality of data. Since $\alpha=\beta<0.05$ serves to constrain the induction

quality of data, RPS test-results associate with comparatively smaller error-rates—particularly a smaller β -error-rate—than a typical published NHST- or BHT-result. Given non-zero error-rates, of course, RPS test-results are not rather than certain and absolute, but probabilistic and relative, as should be.

Given $\alpha=\beta<0.05$, the corroboration quality of hypotheses is sufficient if $LR_{H_i/H_j}>0.95/0.05=19$, thereby comparatively corroborating the point- H_i against the point- H_j . As the error-rates decrease symmetrically as N increases, the value of LR_{H_i/H_j} can be interpreted as the *strength of the empirical warrant* that (among all considered hypotheses) it is the point- H_i , rather than the point- H_j , that matches the population effect size. All pivotal assumptions regarding the error-rates, sample sizes, etc. thus are open and well-motivated, rather than questionably unconstrained or “hidden” in prior probability distributions.

Compared to what $\alpha=\beta<0.05$ demands, of course, the typical sample size in behavioral science today is (far) too small for a fair and firm RPS-hypothesis test. It thus counts towards theoretical progress that, using log-LR addition, individual labs can pool *underpowered* studies into a well-powered global hypothesis test. So, given limited resources, developing an empirically adequate theoretical construct requires that sufficiently large samples are collected cooperatively. For, other than by data, an underpowered population effect size estimate cannot be improved in evidence-based ways.

As a rule, researchers may want to point-specify the H_1 as soon as a small p -value indicates a *preliminary discovery*. Absent trustworthy sources to inform a point-specification, opting for a medium effect size ($d_{H1}=0.50$) can go a long way. Decisively, once the $H_{0(d=0)}$ and the $H_{1(d=x)}$ are point-specified, the corroboration quality of rivaling hypotheses can be evaluated fairly, firmly, and symmetrically, for the H_0 and the H_1 . The so-generated point-effect size estimate is exactly what a theoretician requires to develop an empirically adequate theoretical construct in evidence-based ways. This task presupposes fair and firm test-results, and stands or falls with default point- H_1 -hypotheses.

Declarations/Acknowledgements

FZE and EHW declare that they have no conflict of interest regarding the authorship or the publication of this article. EHW wrote a first draft, which both authors edited iteratively. Both authors approved the final version. FZE acknowledges funding from the Scientific and Technological Research Council of Turkey (project no: 118C257). Thanks to the MTR research group for useful comments that helped improve an earlier draft, especially to Zeynep Burçe Gümüşlü, Irmak Ergin, and Ohan Karatoprak.

References

Anderson, S.F., & Maxwell, S.E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, 21(1), 1–12.

- American Psychological Association (2020). *APA Dictionary of Psychology Online*. <https://dictionary.apa.org/> (accessed 6 April 2020).
- Atneave, F. (1959). *Application of information theory to psychology*. New York: Holt.
- Bakker, M., Dijk, van A., & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554.
- Bandyopadhyay, P.S., & Forster, M.R. (2011). *Handbook of the Philosophy of Science. Vol.: 7 Philosophy of Statistics*. Amsterdam: Elsevier.
- Bayarri, M.J., Benjamin, D.J., Berger, J.O. & Sellke, T.M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, 72, 90–103.
- 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). John Wiley & Sons, Ltd.
- Bird, A. (2020). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*. doi:10.1093/bjps/axy051
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological review*, 111(4), 1061.
- Bryan, C.J., Tipton, E., & Yeager, D.S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behavior* (online first). <https://doi.org/10.1038/s41562-021-01143-3>
- Button, K. S., et al. (2015). Minimal clinically important difference on the Beck Depression Inventory – II according to the patient’s perspective. *Psychological Medicine*, 45, 3269-3279.
- 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>
- Cohen, J. (1977, rev. ed). *Statistical power analysis for the behavioral sciences*. London: Academic Press.
- Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, 16(4), 779-788 <https://doi.org/10.1177/1745691620970586>
- Eronen, M. I., & Romeijn, J. W. (2020). Philosophy of science and the formalization of psychological theory. *Theory & Psychology*, 30(6), 786-799.
- Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE* 11(2), 0149794.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- 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.

- Fienberg, S. E. (2006). When did Bayesian inference become “Bayesian”? *Bayesian analysis*, 1(1), 1-40.
- Fisher, R.A. (1925, 1972). *Statistical methods for research workers*. London: Oliver & Boyd.
- Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, 7, 585-594.
- Francis, G. (2014). The frequency of excess success for articles in Psychological Science. *Psychonomic Bulletin & Review*, 21, 1180-1187.
- Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: publication bias and psychological science’s aversion to the null. *Perspectives on Psychological Science*, 7, 555–561. doi: 10.1177/1745691612459059
- Feynman, R. P. (1974). Cargo cult science. *Engineering and Science*, 37(7), 10-13.
- Gervais, W.M., et al. (2015). A powerful nudge? Presenting calculable consequences of underpowered research shifts incentives toward adequately powered designs. *Social and Personality Science*, 6(7), 847-854.
- Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology* 8 (2), 195–204.
- 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.org/10.1177/2515245918771329>
- Greenland, S., Senn, S.J., Rothman, K.J. et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31, 337–350. <https://doi.org/10.1007/s10654-016-0149-3>
- Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... , & Calvillo, D. P. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573.
- Hempel, C. G., and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135-175.
- Irvine, E. (2021). The role of replication studies in theory building. *Perspectives on Psychological Science*, 16(4), 844-853. <https://doi.org/10.1177/1745691620970558>
- Ioannidis, J.P.A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, 686-701.
- Jaynes, E.T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4), 620-630.
- Jeffreys, H. (1931). *Scientific inference*. Cambridge: Cambridge University Press.
- Jeffreys, H. (1948, 1961³). *Theory of probability*. Oxford: Oxford University Press.
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., ..., & Nosek, B. A. (2018, November 21). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting. *Advances in Methods*

- and Practices in Psychological Science*, 1(4), 443-490.
<https://doi.org/10.17605/OSF.IO/8CD4R>
- Krefeld-Schwalb, A., Witte, E. H., & Zenker, F. (2018). Hypothesis-testing demands trustworthy data—a simulation approach to statistical inference advocating the research program strategy. *Frontiers in Psychology*, 9, 460.
- Lakatos, I. (1978). *The Methodology of Scientific Research Programs, Vol. 1*. Cambridge: Cambridge University Press.
- Lakens, D., Scheel, A.M., & Isager, P.M. (2018). Equivalence testing for psychological research: A Tutorial. *Advances in Methods and Practices in Psychological Science*. 1(2), 259-269.
- Linden, A. H., & Hönekopp, J. (2020). Heterogeneity of research results: a new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*,
<https://doi.org/10.1177/1745691620964193>
- Locascio, J.J. (2017). Results blind science publishing. *Basic and Applied Social Psychology*, 39, 239-246.
- Locascio, J.J. (2019). The impact of results blind science publishing on statistical consultation and collaboration. *The American Statistician*, 73:sup1, 346-351.
- Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, 72, 19–32.
- Mayo, D.G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.
- Mayo, D.G., & Spanos, A. (2011). Error statistics. In: P.S. Bandyopadhyay & M.R. Forster (Eds.) *Handbook of Philosophy of Science. Vol. 7: Philosophy of Statistics* (pp. 153–198). Amsterdam: Elsevier.
- Maxwell, S.E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, 9 (2), 147–163.
- Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.
- Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108-141.
- Miłkowski, M.; Hohol, M., & Nowakowski, P. (2019). Mechanisms in psychology: The road towards unity? *Theory & Psychology*, 29(5), 567-578.
- Morey, R.D., Romeijn, J.-W., & Rouder, J.N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6–18.
- Neyman, J., & Pearson, E. S. (1967). *Joint Statistical Papers*. Cambridge: University Press.

- Norman, G. R., Sloan, J. A., & Wyrwich, K.W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, 41, 582-592.
- Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, 26(5), 1596-1618.
- Olsson-Collentine, A., Wicherts, J.M. & van Assen, M.A.L.M. (2020, July 23). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*. Advance online publication. <http://dx.doi.org/10.1037/bul0000294>.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, 6251, acc4716-1-8.
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on Psychological Science*, 7(6), 528-530.
- Popper, K.R. (1959). *Logic of discovery*. London: Routledge
- Reichenbach, H. (1938). *Experience and Prediction*. Chicago, IL: University of Chicago Press.
- Richard, F.D., Bond, C.F., Jr., & Stokes-Zoota, J.J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331-363.
- Robinaugh, D., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. (2020, March 25). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. <https://doi.org/10.31234/osf.io/ugz7y>
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rothstein, H.R. Sutton A.J. & Borenstein M. (2005). *Publication Bias in Meta-Analysis – Prevention, Assessment and Adjustments*. John Wiley & Sons.
- Royall, R. (1997). *Statistical evidence. A likelihood paradigm*. London: Chapman & Hall.
- Schauer, J.M. & Hedges, L.V. (2020, April, 9). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*. Advance online publication. <http://dx.doi.org/10.1037/bul0000232>.
- 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>
- Schönbrodt, F., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25, 128-142. <http://dx.doi.org/10.3758/s13423-017-1230-y>
- Stanley T. D., Carter, E. C., & Doucouliagos, H. (2018). What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin*, Vol. 144 (12), 1325–1346. <http://dx.doi.org/10.1037/bul0000169>
- van Rooij, I., & Baggio, G. (2020). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives*

- on *Psychological Science*, 16(4), 682-697.
<https://doi.org/10.1177/1745691620970604>
- Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143, 1457–1475.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H.L.J., & Kievit, R.A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7, 632–638.
- Wald, A. (1947). *Sequential Analysis*. New York: Wiley.
- Witte, E.H., and Zenker, F. (2016). Reconstructing recent work on macro-social stress as a research program. *Basic and Applied Social Psychology*, 38(6), 301–307.
- Witte, E.H., and Zenker, F. (2017a). Extending a multilab preregistered replication of the ego-depletion effect to a research program. *Basic and Applied Social Psychology*, 39 (1), 74-80.
- Witte, E.H., and Zenker, F. (2017b). 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., and Zenker, F. (2018). Data replication matters, replicated hypothesis-corroboration counts. (Commentary on “Making Replication Mainstream” by Rolf A. Zwaan, Alexander Etz, Richard E., Lucas, and M. Brent Donnellan.) *Behavioral and Brain Sciences*, 41, e120.

Total word count: 12861