The statistical, theoretical, and practical aspect of an empirical effect—On the smallest effect of interest for theory construction in behavioral science

Erich H. Witte and Frank Zenker*

*correspondence: fzenker@gmail.com

Abstract: A transparent evaluation of the scientific relevance of an observed empirical effect should distinguish between the size of effect (*statistical aspect*), the ability of a theoretical construct to predict the effect (*theoretical aspect*), and the utility associated with it (*practical aspect*). The ‘smallest effect of interest’ (SEOI) for an empirically adequate explanation is for formal reasons smaller than the SEOI for an empirically adequate prediction. As recent meta-meta-analyses show, homogeneous observed effects in behavioral science are typically small, or otherwise highly heterogeneous. The SEOI for an empirically adequate prediction is therefore typically larger than what can be observed. As this poses a severe challenge to theory construction research, it (partially) explains why empirical adequate theoretical constructs here are rare.

Keywords: effect size; error-rates; explanation; prediction; theory evaluation; utility

1. Introduction

‘Effect size’ denotes the amount of influence or correlation between independent and dependent variables, as observed or predicted under experimental or natural conditions. Behavioral science publications typically report observed effects using standardized effect size measures—such as d (Cohen 1977) or the related measure r (Pearson, 1895; see Schäfer & Schwarz, 2019)—and conventionally define effects as *small*, *medium*, or *large* ($d=0.20$, $d=0.50$, $d=0.80$; Cohen, 1977, Table 2.2.1). Whether an observed effect is reported as a correlation (r), a determination coefficient (r^2), or a relative increase in empirically adequate prediction, Rosenthal and Rubin (1982, 166) rightly observed that

“[...] experienced behavioral researchers and experienced statisticians [reacted] quite surprised when we showed them that the Pearson r of .32 [or $d=0.65$] associated with a coefficient of determination (r^2) of only .10 was the

correlational equivalent of increasing a [treatment's] success rate from 34% to 66% by means of an experimental treatment procedure.”

A medium observed effect of $r=0.32$ thus goes along with a noteworthy improvement in a treatment's success rate. But noteworthiness says nothing about the size of effect that is relevant for theory construction. We will argue that an effect of size x is relevant for theory construction only if its stable observation can support a theoretical construct that predicts x .

As a review of published effects in behavioral science would show, the evaluation of observed effects here typically oscillates between the effect's size and its utility. An instructive example is how Greenwald, Banaji, and Nosek's (2015) reply to Oswald, Mitchell, Blanton, Jaccard & Tetlock's (2013) meta-analysis of the *Implicit Association Test* (IAT). Their meta-analysis showed that IAT scores correlate to a mere $r=0.148$ with the racial/ethnically discriminatory behavior the IAT tests for. In fact, Greenwald et al.'s (2015) own meta-analysis also reports a *small* effect, ($r=.22$ or $d=0.44$). But the authors nevertheless defend the IAT by invoking its utility:

“Oswald, Mitchell, Blanton, Jaccard, and Tetlock concluded that IAT measures show ‘poor prediction of racial and ethnic discrimination’ (pp. 171, 183) and provide ‘little insight into who will discriminate against whom’ (p. 188). [But] Oswald, Mitchell, Blanton, Jaccard, and Tetlock's (2013) conclusion did not take into account that small effect sizes *affecting many people* or *affecting individual people repeatedly* can have great societal significance.” (Greenwald et al., 2015, 560; *italics added*, notation adapted)

Greenwald et al. (2015) thus shift the issue to the IAT's practical relevance. Since even the smallest (non-zero) effect retains practical relevance if it applies often in a population, this warrants a positive evaluation. Oswald et al.'s (2013) negative evaluation instead cited that $r=0.22$ implies that “[t]he low *predictive* utility for the race and ethnicity IATs present problems for contemporary theories of prejudice and discrimination that assign a central role to implicit constructs” (Oswald et al., 2013, 183). What speaks against the IAT's *theoretical* relevance, then, is that it fails to

adequately *predict* observations. So while both teams of authors use the same statistical and meta-analytical methods, they nevertheless use different evaluative aspects.

A transparent evaluation arguably requires a *unique* evaluative aspect. The one-dimensional evaluation suggested by Cohen's (1977) *d*-measure fails this requirement. Other related effect size measures do the same (e.g., partial *r*, η^2 , odds-ratio; Ferguson, 2009). A transparent evaluation, we submit, must distinguish three aspects: the effect's size (*statistical aspect*), how well a theoretical construct subsumes it (*theoretical aspect*), and the effect's utility (*practical aspect*). Under the statistical aspect, one does not so much evaluate but merely *describe* the observed effect. This description widely counts as an *explanation*. Under the theoretical aspect, moreover, one evaluates the agreement between an observed effect and a theoretical prediction. And under the *practical* aspect, one evaluates the effect's utility.

The practical aspect is independent of the statistical and theoretical aspects. Thus, even a very small effect with non-zero utility can be practically relevant. A very small effect, however, typically lacks theoretical relevance. The reason is that two distinct error-pairs apply to quantify the uncertainty when a theoretical construct *explains* an observed effect (α -, β -error-rates), as opposed to when a construct *predicts* an effect that is yet-to-be observed (γ -, δ -error-rates). The 'smallest effect of interest' (SEOI) for an empirically adequate prediction, therefore, is *larger* than the SEOI for an empirically adequate explanation.

Several recent meta-meta-analyses of published effects in behavioral science show that observed effects are typically small and homogenous or large(r) and highly heterogeneous. The latter thus entail large observed variance (Schauer & Hedges, 2020; Olsson-Collentine, Wicherts & van Assen, 2020; Linden & Hönekopp, 2021). The homogenous share of observations therefore implies that these observations are overlain by the standard measurement error of 5%, becoming quasi-unobservable. And the heterogenous share of observations implies that predicting such observations incurs unacceptably large error-rates. The upshot is that an *empirically adequate* theoretical construct would have to predict a much larger, homogeneous effect than can be observed. Since this impairs theory construction research in behavioral science, it can (partially) explain the rarity of empirical adequate theoretical constructs here.

We briefly introduce the three aspects (Sect. 2), then develop each (Sects. 3-5), and conclude (Sect. 6).

2. Statistical, theoretical, and practical aspect

2.1 *The smallest effect of interest for theory construction*

Empirically, the term ‘effect size’ refers to *any* observed mean difference regardless of its size (Cohen, 1977). To be a *theoretically worthwhile* parameter, by contrast, the mean difference must justify the efforts of subsuming it under a theoretical construct. Given the goals of a cumulative science of human behavior, defining a SEOI for a theoretically worthwhile parameter could provide a useful convention, comparable to Cohen’s (1977) error-rates for a statistical test ($\alpha=0.05$; $\beta=0.20$). As criteria for *theoretically worthwhile* SEOI, we propose (i) that the effect must be observed under error-rates of $\alpha=\beta<0.05$ in a sample of at least $N_1=N_2=53$, and (ii) that the effect explains $\geq 10\%$ of the observed variance. (The observed variance is a squared correlation with the experimental grouping factor.)

Together both criteria yield an SEOI of $d \geq 0.65$, a compromise between a *medium* ($d=0.50$) and a *large* effect ($d=0.80$). Provided this effect is *stably* observed (by meeting rigorous thresholds of statistical significance and statistical power relative to the above minimum sample size), $d \geq 0.65$ would offer a theoretically motivated constraint on what publications *should* report to be relevant for theory construction research. In Rosenthal and Rubin’s (1982) case (Sect. 1), for instance, $r=0.32$ ($d=0.65$) is statistically significant in a sample of $N_1=N_2=100$ and explains $\geq 10\%$ of the observed variance. Even the smallest proportion of explained observed variance, of course, can be “excused” if a very small non-zero observed effect with non-zero utility applies often in a population. Utility considerations alone therefore warrant a *much smaller* SEOI than what is required for a theoretically worthwhile parameter.

That said, the main issue for theory construction research is that behavioral science publications typically report *small to medium* observed effects ($0.20 \leq d \leq 0.60$; Bakker, et al., 2012). Among these only small effects are observed with a small enough observed variance to be deemed *homogeneous* (Schauer & Hedges, 2020; Olsson-Collentine, Wicherts & van Assen, 2020; Linden & Hönekopp, 2021). Moreover, critical analyses of replications studies and meta-analytical methods strongly suggest

that meta-analytical effect size estimates are vast *overestimates* (Klein et al., 2018; Schäfer, 2018). At both object- and meta-level, therefore, published observed effects would seem to *fail* as worthwhile theoretical parameters.

We return to this below (Sect. 3). Let us first turn to the statistical aspect.

2.2 The statistical aspect

Under the statistical aspect, an observed effect is described as a parameter estimate. This estimate requires the use of a statistical significance measure and an effect size measure. The classical measure of null-hypothesis significance testing (NHST), the p -value (Fisher, 1959) or the α -error (Neyman & Pearson, 1967), states the probability (P) of data (D) in view of a null-hypothesis (H_0), $0 \leq P(D, H_0) \leq 1$.¹ Conventionally, if $P(D, H_0) \leq \alpha$, then the effect is said to be statistically significant (Fisher, 1959). But this says nothing about the *difference* between D and H_0 —i.e., the absolute size of observed effect, ‘ $|d| = (m_1 - m_0)/s$ ’ (where $0 \leq |d| \leq \infty$), as a “specific non-zero value in [a] population” (Cohen, 1977, 10)—because the d -measure varies independently of the significance level (p , α) and the sample size (N). (Though inverse effects are conventionally stated on the negative arm of the number line, there are no meaningful “negative effects.”) Given a fixed sample size, therefore, the p -value, respectively the α -error, is *complementary* to the d -value.

Obtaining a statistically significant effect hence depends not directly on d , but rather depends on N . Indeed, sufficiently increasing N lets even the virtually smallest possible effect become statistically significant eventually (Bakan, 1966). For instance, if a t -test establishes $P(H_0, D)$ as statistically significant, then the numerator of in $t = (m_1 - m_2) / \sqrt{(s_1^2/N_1 + s_2^2/N_2)}$ states the observed mean difference ($m_1 - m_2$) in experimental and control groups (N_1 , N_2), and the denominator states the standard error associated to the means. Given $(m_1 - m_2) = 0$, we find $t = 0$, independently of N . Whereas if $(m_1 - m_2) \neq 0$, then

¹ The statistical approaches of Fisher (1959) and Neyman-Pearson (1967) differ epistemologically. Fisher treats the p -value as a measure of empirical evidence that is inductively inferred from data. Neyman-Pearson instead treat the α -error-rate as a measure of the long-run chances that the same inductive inference is erroneous. In praxis, alas, that difference often “disappears” (Hubbard & Bayarri, 2008)

a sample size that lets $(m_1 - m_2)$ pass *any* threshold of t can always be found. Of course, the t -value increases with the numerator. But if $m_1 - m_2$ is held constant, then the t -value increases *only* with the denominator. So, increasing $N_{1,2}$ is the only way of increasing t .

The t -value thus varies with the non-theoretical parameter N . Yet N pertains to a statistical estimation, but not to a theoretical explanation. The p -value hence determines whether an observed effect is *statistically significant*. Determining an effect's *scientific relevance* additionally requires "provid[ing] some effect size estimate when reporting a p -value" (Wilkinson & APA Task Force, 1999, 599; notation adapted; see Huberty, 2002). Although effect size estimates are formally required for a statistical power-analysis, only some 5% of published studies in behavioral science are estimated to conduct this analysis (Schäfer, 2018; Schäfer & Schwartz, 2019; Stanley, Carter, & Doucouliagos, 2018). Behavioral science publications do thus seem to *avoid* effect size estimates, whether they be motivated empirically or theoretically.

2.3 The theoretical aspect

If 'effect size' refers to an *observed* difference between a point- H_0 and a point- H_1 , one presupposes an inductive interpretation (e.g., when estimating the influence of macro-social stress on physically experienced pain; Witte & Zenker, 2016). When 'effect size' refers to a *predicted* difference—a usage that presupposes a deductive interpretation because the observed difference is derived from a theory—two crucial inductive assumptions nevertheless remain in play. They concern the observed variance and the measurement error.

As for the *observed variance*, experience teaches that observations vary around the mean. Researchers must therefore center a first data distribution on the mean stated by what typically is a *random* version of H_0 and must center a second distribution on the mean stated by a *substantial* H_1 . In most cases, however, a theoretical construct can only predict the mean difference $(m_1 - m_0)$, but not the true variance of behavioral responses (aka 'person variance' in differential psychology). Rather than deducing the two distributions from a theory, therefore, researchers must typically induce them from observations. As for the *measurement error*, moreover, experience also teaches that empirical observation is error prone. The measurement error for observed behavioral

responses, however, is what a theoretical construct cannot provide either. So it must again be specified inductively.

Since the *validity* of a theoretical construct can be interpreted as the proportion of data that a construct adequately subsumes, or “saves” (see Duhem, 2015), both distributions—which as inductive parameters represent a combination of the person variance and the measurement error—do affect a construct’s validity. Both inductive parameters thus co-determine the theoretically predicted effect size. The latter must therefore be stated *independently* of inductive parameters. This means the theoretically predicted effect must be specified *prior* to data collection, and assumptions about expected measurements must be explicated. (We return to this in Sect. 4.)

2.4 The practical aspect

A key-example of purely inductive research is the meta-analytical estimation and aggregation of 2nd- and 3rd-order object-level observed effects by means of data-based statistical assumptions, e.g., fixed or random effect models (Grissom & Kim, 2012; Schulze, 2004; Cafri, Kromrey & Brannick, 2010; Mitchell, 2012; Richard et al., 2003; Schmidt, Oh & Hayes, 2009). 1st-order effects are observed study results, 2nd-order effects arise in a discipline’s sub-areas (e.g., attitude research *vs.* small group research), and 3rd-order effects in sub-disciplines (e.g., social *vs.* cognitive psychology).

A major difficulty is that *publication bias* causes a meta-analytical effect size estimate to vastly over-shoot a true effect (Klein et al., 2018, Schäfer, 2018, Stanley et al., 2018). In behavioral science, moreover, object-level observed effects arise in qualitatively different domains and are normally rather small (Stanley, Carter & Doucouliagos, 2018). Klein et al.’s (2018) replication study suggests that the *median* observed effect size would generally reduce from $d=0.60$ in an original study to $d=0.15$ in a replication study. This is far smaller than what Cohen (1977, 22) defined as a *medium* effect ($d=0.50$). For some 100 years of social psychological research, a 3rd-order analysis even estimates the average observed effect as a mere $r=0.21$, or $d=0.42$ (Richard et al., 2003). For general psychology we may assume $r=0.16$, or $d=0.33$ (Cafri, Kromrey, & Brannick, 2010).

Were three independent researchers to evaluate $d=0.33$, the first might find it “normal,” because it falls within the typical range of $0.20 \leq d \leq 0.60$ (Bakker, et al.,

2012). A second researcher instead may evaluate $d=0.33$ as an *unacceptable* correlation between a theoretical construct and an empirical indicator (i.e., a test-validity coefficient), because $d=0.33$ explains $<10\%$ of the observed variance (Cohen 1977, 22; Table 2.2.1).² Behavioral predictions with attitude models, by contrast, predict on average $r=0.40$, or $d=0.90$, while explaining some 16% of variance (Eckes & Six, 1994).) The third researcher, finally, may find $d=0.33$ too small for theory construction, because observing $d=0.33$ (even repeatedly) is insufficient to inductively support a construct that predicts $d=0.33$. Because as the effect size decreases, uncontrolled factors contribute ever more to it. Yet these factors are precisely what one does *not* model theoretically.

If $d=0.33$ results from an intervention with positive consequences, however (e.g., a reduction of the risk of a heart attack; Rosenthal, 1990), then all three researchers could agree on its *practical* relevance. Indeed, an “intervention with [a] weak effect size but no risks may be valuable” (Ferguson, 2009, 536). Given “big data,” a very small effect for an individual can turn into a vast social effect. Optimizing the match between a commercial product and consumer preferences, for instance, reportedly correlates with a 1% increase in life-satisfaction for 244 million consumers. This is how Matz, Gladstone & Stillwell (2017) reply to Boyce et al.’s (2017) objection that a 1%-increase hardly matters. Analogous to how Greenwald et al. (2015) reply to the objection that the IAT’s small effect ($r=0.148$; $d=0.30$) its measurement procedure invalid. Similar replies *must* cite an effect’s utility, invoking the practical aspect.

² Since an experimental condition can be understood as the empirical *realization* of a theoretical construct, the experimental condition can be likened to a standardized test of a psychological variable, a test that should be validated before using it. How a specific experimental condition is realized (including the setting, the task instructions, and how participants interpret them) thus is logically equivalent to how a standardized test measures a theoretical construct (e.g., attitude, intelligence, personality). Generally, a test can be treated as measuring a psychological construct, only if individual behavioral responses sufficiently correlate with an independent criterion (e.g., an expert rating). This is most obvious for binary ratings (e.g., true/false judgements), but extends fully to graded ratings (Borsboom, Mellenbergh & van Heerden, 2004).

We proceed to treat each aspect in detail.

3. The statistical aspect

To explain observations, the probability of observed data, as well as of more extreme data, must satisfy the statistical significance criterion $P(D, H_0) < p$, or $P(D, H_0) < \alpha$. As we saw, evaluating the statistical aspect of an effect requires an *inductive* inference of a parameter estimate from observations. To establish this estimate as a statistically significant difference in the dependent variable, empirical studies should ideally use two samples (N_1, N_2). Even without an actual control group, the observed mean in the treatment group (m_1) can always be compared to a random effect. Moreover, one can in principle proceed from a *sample* parameter estimate to a *population* parameter estimate, namely by modifying the d -measure (e.g., via least square or maximum likelihood estimates), (Lakens, 2013; Grissom & Kim, 2012; Schulze, 2004).

Already the sample estimate, of course, depends on a study's design choices—e.g., whether one or several means are used (as in ANOVA or structural, hierarchical equation models)—as well as on the quality of the measurement scale (nominal, ordinal, interval, ratio).³ Regardless, $d = (m_1 - m_2)/s$ invariably relates the observed mean difference ($m_1 - m_2$) in an individual study explicitly to the observed standard deviation of behavioral responses (s_{obs}) and implicitly to the statistical measurement error ($s_{\text{error}}/\sqrt{N-1}$) associated to both means. For this reason, Cohen's d -measure ever incurs the need to handle the observed s . This need leads to distinct consequences when *explaining* an observed effect as opposed to *predicting* it.

These consequences particularly matter for a meta-analytical aggregation of object-level effects into a global effect. This requires a neutral d -measure, where d is independent of N and the measurement scale. A neutral d -measure seems to be easily available, because the d -measure invariably compares observations with a point-

³ Using the d -measure to estimate an effect size on a *nominal* or an *ordinal* scale, for instance, differs from using it on an *interval* scale. Nominal and ordinal scales only allow for discrete measurement, with interval scales allow for continuous measurement. As statistical test in behavioral science tend to use normally distributed approximations, one thus behaves *as if* the observed parameter is measured continuously (Cohen, 1977).

hypothesis stating the mean of a random normal distribution ($d_{H0}=0$) and one stating the mean of an experimental group distribution ($d_{H1}=x$). The expression $d_{H1}=x$ thus implies that (m_1-m_0) states the observed effect *as standardized* by the observed s , or in a meta-analytical aggregation as standardized by the pooled s . So, if s takes the same value in both groups (N_1, N_2), then (m_1-m_0) states d neutrally. The observed effect thus rests inductively on observations alone.

Meta-analyses and particularly recent meta-meta-analyses, however, strongly suggests that object-level studies in behavioral science typically report homogeneous effects that are too small to be observed—because the standard measurement error of 5% overlays them—or otherwise typically report medium to large effect sizes that are highly heterogeneous (Schauer & Hedges, 2020; Olsson-Collentine, Wicherts & van Assen, 2020; Linden & Hönekopp, 2021). Heterogeneity means that the observed d -values of aggregated object-level studies vary greatly because observed behavioral responses display large variance. The heterogeneity of observations necessarily inflates the uncertainty associated with the process theoretically postulated to subsume observations. This suggests that experimental settings are subject to *qualitatively different* influences (Stanley et al., 2018; Viechtbauer, 2007). Although a moderator analysis may offer cues how to control for heterogeneity, these analyses rarely identify promising moderators (Linden & Hönekopp, 2021). And the influence of moderators must always be modelled theoretically. This, however, rarely occurs. Some researchers hence reject a parameter estimate based on heterogeneous measurements (Sixtl, 1993).

Since the d -value varies independently of theoretical assumptions about (m_1-m_0) with the variation of s , one can use s as a new (implicit) measurement unit. Observations are thus treated *as if* they are measured on a scale whose measurement differs from the unit of an original measurement scale (Baguley, 2009). A Likert-scale, for instance, retains two free parameters: the origin and the unit. Because differences are invariant under addition of a constant (that shifts the scale's origin), the origin can be set to (m_1-m_0) . And, regarding the unit, measurement-theory rightly assumes that individual behavioral responses are normally distributed around the mean. Different Likert-scales can therefore yield different values of s , wherefore the effect's *statistical* meaning comes to depend on s . Hence, in statistical testing the measure of variability in a sample is not (m_1-m_0) itself, but rather (m_1-m_0) as standardized by s .

How $(m_1 - m_0)$ varies in a sample thus depends on its numerical value, as well as on the observed standard deviation relative to the measurement scale. The main reason why s regularly serves as a measurement-unit is that behavioral science research lacks measurement scales with a *theoretically motivated* measurement unit (Flake & Fried, 2020). When conducting statistical descriptions or tests, therefore, it regularly proves impossible to get rid of the inductive measurement unit that s is. Under the statistical aspect, therefore, the term ‘effect size’ denotes an *inductive parameter estimate*—a combination of the observed $(m_1 - m_0)$ and the observed s . Via a *normalization* (that sets the original scale’s origin to 0) and a *standardization* (that eliminates the variability of s by setting $s=1$), this parameter estimate uses the same standardized measurement scale for all object-level studies. Consequently, d -values turn into standard z -scores, which are directly comparable across two or more different studies. This, however, is paid for by focusing on mean-differences, and by using the inductive parameter s .

The theoretical aspect instead addresses only $(m_1 - m_0)$, “liberating” it from s , as it were.

4. The theoretical aspect

4.1 A parameter for well-planned research

The all-things-considered best measurement scale, of course, is simply unknown. But although the use of different measurement units for a Likert-scale may yield different standard deviations, it certainly cannot yield different deviations of means. For this very reason, one can normalize s in the first place, by setting $s=1$. Consequently, with $(m_1 - m_0)/s$ now expressed on a scale with the neutral measurement unit $s=1$, one retains $(m_1 - m_0)$ as a *non-inductive* parameter. This parameter states the theoretical effect size as the difference between *average* observed behavioral responses.

Interval scale measurements, as we saw, leave the *origin* and the *unit* of a scale undetermined. As we also saw, neither parameter can be determined theoretically if the measurement of a variable lacks a theoretically motivated measurement unit and origin that are *certain*. This uncertainty is precisely what a standardization eliminates. For instance, standardizing measurements in experimental and control group to the grant mean $M=0$ retains the difference to the control group’s mean, $X_{\text{control}} - M$, on one hand, and the difference to the experimental group’s mean, $X_{\text{exper}} - M$, on the other. Since M

cancels out, $X_{\text{control}} - X_{\text{exper}}$ becomes a standardized difference measurement. This is exactly how raw measurement scores are transformed into z -values: standardize the control group's mean to $m_0=0$, the experimental group's mean to $m_1=d$, and the standard deviation in both groups to $s=1$. In brief, this describes the assumption that z -values are standard normal curve-distributed, or t -distributed.

This assumption suffices to separate the theoretically predicted parameter ($m_1 - m_0$) from the inductive parameter s , because both parameters become *independent* given a normal distribution. Whether the theoretically predicted effect then grounds in observations, reconstructed empirical results, or a theory's formal core, this assumption warrants interpreting ($m_1 - m_0$) as the *theoretically expected* influence of independent onto dependent variables, respectively the correlation between them (Witte, 1996). Consequently, Neyman-Pearson test-theory (NPTT) can determine the *minimum sample size* (N_{MIN}) for a well-powered study to register the theoretically predicted ($m_1 - m_0$) as a statistically significant effect. So, although most published behavioral science studies are in fact underpowered (Bakker et al. 2012; Stanley et al., 2018) and while most behavioral science theories today cannot predict a point-specific ($m_1 - m_0$), a point-specific effect is an indispensable parameter to plan sound empirical research (Cohen, 1977; Witte & Zenker, 2017).

4.2 Theoretical and empirical effect size

4.2.1 Explanation vs. prediction

An empirically adequate theoretical construct must not only *explain* an effect that is observed under controlled laboratory conditions. It should also *predict* a similar effect under natural conditions. However, the fewer states of affairs a vaguely predictive construct prohibits, the larger is the class of *theoretically non-modelled* influences on an experimental setting that potentially explain the observed effect. Leaving a theory vague thus turns all non-modelled influences into potential explanatory factors for a non-random effect. Conversely, a theoretically embedded point-specific effect warrants excluding all those candidate explanatory factors that cannot produce the effect.

To evaluate a theoretical construct's ability to predict and explain data, the most useful tools probably are the error-rates themselves. After all, given a constant quality of measurement and constant (interval scale-based) standard errors associated to m_1 and

m_0 , as the effect size increases, the normal distribution under H_0 ($m_0=0$) overlaps ever less with the normal distribution under H_1 ($m_1=d_x$). As data thus become more probable in view of H_1 than in view of H_0 , i.e., $P(D, H_1) > P(D, H_0)$, a theoretical construct stating H_1 predicts a larger proportion of data adequately. So, the larger the predicted effect, the better can a theoretical construct separate a systematic from a random effect. Hence, *ceteris paribus*, theoretical constructs stating *larger* effects are better predictive tools.

The following two examples illustrate this. (Because nothing depends on their substance, readers may easily construct their own examples.)

4.2.2 Examples

Classical logic is binary. It readily models whether a measurement instance indicates a property's presence or its absence (nominal scale measurement). A case in point is biological sex (rather than gender). If genetic abnormalities are not assumed, then all men, and only all men, possess the Y-chromosome. This means the chromosome's effect on sex does project fully onto the male/female-distinction. Except for the measurement error, a someone's biological sex can thus be predicted *perfectly*, i.e., the effect size is 100%. Presence of the Y-chromosome also correlates with body height: a Y-chromosome bearer is *on average* taller than a non-Y-chromosome bearer. Some females are nevertheless taller than the average male. Mapping the properties 'male' and 'female' onto length-scale intervals thus results in an *imperfect* prediction of sex from height. The uncertainty is $d = (m_1 - m_2) / s = 183\text{cm} - 166\text{cm} / 10\text{cm} = 1.7$. A clothing manufacturer might find this *very large* effect a good rule-of-thumb to purchase fabrics.

In behavioral science, by contrast, very large effects are a rarity. One such rarity is the estimated degree of linear relationship between mental abilities (measured as an IQ-score) and a successful performance on a moderately demanding job. Schmidt and Hunter's (1998) meta-analysis estimates the effect as an correlation of $r=0.51$ or $d=1.2$. This too is a *very large* effect, and no single other measure predicts job performance better. An employer who would exclusively hire job-applicants whose intelligence scores exceed the critical value for $d=1.2$ given symmetrical errors ($\gamma=\delta=0.05$)—namely $z=0.60$ —should thus expect 73% of correct predictions (see Fig. 1) and to retain a prediction-error of $\gamma=\delta=100\%-73\%=27\%$. As the effect shrinks, however, the prediction-error grows larger. In case of $z=0.40$ ($d=0.80$), for instance, the prediction-

error increases to 34%. Generally, the smaller the predicted effect is, the less certain is the prediction.

4.2.3 Explanation vs. prediction error-rates

Using observed data to evaluate the empirical adequacy of a theoretical construct is *conceptually* different from using an empirically adequate construct to predict new data. Our hiring example, for instance, predicted nothing. It merely used two normal distributions, centered on H_0 and H_1 , to transform the data-based parameter $d=z=1.2$ into a measurement model for a hiring decision. But this does not allow to evaluate whether $d=z=1.2$ as an (assumedly true) *empirical* parameter confirms $d=z=1.2$ as a (hypothetical) *theoretical* parameter.

Modelling old observations and predicting new observations is associated with qualitatively distinct error-rates, that answer to different needs. The γ - and δ -error-rates (*prediction*) differ from NPTT's α - and β -error-rates (*explanation*). The former allow to estimate the proportion of erroneous predictions when using a theoretical construct that states a true parameter. As Fig. 1 shows, the γ - and δ - error-rates vary only with the z -value, itself an element of the assumedly true theoretical parameter. NPTT's α - and β -error-rates, by contrast, denote the proportion of data leading to a *mistaken* acceptance-/rejection-decision regarding a theoretical parameter that *stands proxy* for an observed parameter.

Each pair of error-rates varies with the theoretical parameter (the z -value). But the α - and β -error-rates additionally vary with N . Thus, the degree of error-proneness associated to a decision also varies with N . Hence, if N is constant, then the α - and β -error-rates increase as the theoretical parameter decreases. A decrease in the theoretical parameter, therefore, must be compensated by an increase in N . Only then is a constant degree of error-proneness maintained (i.e., constant α - and β -error-rates).

Fig. 1: Illustration of $z=m_1-m_0=1.2$; $s=1$; $\gamma=\delta=0.27$; $(1-\delta)=0.73$. Here, m_1, m_0 are the means of the areas under the solid and the dashed curve; the abscissa (x -axis) plots possible intelligent scores as z -values; the ordinate plots the normal distribution's probability density; γ and δ are the error probabilities given the z -values.

If the effect a theoretical construct predicts is evaluated statistically under non-zero error-rates, then each measurement instance retains a non-zero probability of constituting an element of the H_0 -distribution ($m_0=0$) *or* the H_1 -distribution ($m_0=1.2$) (Fig. 1). At the mid-point of m_1-m_0 , where both distributions intersect—i.e., at the critical value $z=0.6$ —a measurement instance thus is equally probable as an element of *this* or *that* distribution (50/50 chances). By contrast, $z<0.6$ makes each measurement instance more probable under the H_0 - than under the H_1 -distribution, and *vice versa* for $z>0.6$. Regarding the correlation of variables (e.g., the intelligence score with job performance), therefore, the proportion of data that a theoretical construct predicts adequately is a function of the expected (true) population parameter.

In this sense, the observed parameter states a theoretical prediction's "graded truth." Specifically, $1-\delta=0.73$ implies that 27% of all measurements to both sides of critical z are elements of the *other* distribution. This entails *erroneously* predicting that about ¼-th of all applicants succeed on the job; and the same error-rate applies when poor job performance is predicted. Thus, $z=0.60$ ($d=1.2$) leads to about ¼ of "bad hires" while missing about ¼ of "good hires." By parity of reasoning, if $d=1.2$ decreases to $d=0.80$ (i.e., $z=0.40$), then 66% of hired applicants would presumably enjoy job success, while 34% of (un)successful applicants are erroneously (not) hired (Fig. 2).

Fig. 2: Illustration of $m_1 - m_0 = 0.80$; $s = 1$; $(1 - \delta) = 0.66$. The difference between proportions is based on the means of a normal distribution (see Cohen, 1977, 180); compare Fig. 1.

So even if the observed effect is *large*—and this is rare in behavioral science—the predictive instrument that a good theoretical construct *should* be, can outperform what a fair coin-toss predicts (50/100 cases) in only 16/100 cases. When empirical data are used to *explain* $d = 1.2$, therefore, collecting enough data to meet $\alpha = \beta \leq 0.05$ merely requires sufficient resources. The use of a theoretical construct to *predict* $d = 1.2$, by contrast, incurs so much theoretically unintended (but unavoidable) variability that a decision based on this construct would improve a chance-based decision by only 16%.

4.2.4 Prediction counts as much as explanation

True, “[o]ur goal as scientists is not to publish as many articles as we can, but to discover and disseminate truth” (Simmons, Nelson & Simonsohn, 2011, 1365). Meeting this goal, however, squarely depends on the size of effect. Particularly the SEOI for a theoretically worthwhile parameter depends on the ability of a theoretical construct to explain *and* predict data. Yet, today, as behavioral science predominantly seeks to “discover” true effects, the focus is the use of a theoretical construct to *explain* the behavior of the *average* participant. Whereas if a theoretical construct *predicts* data, the focus shifts to predicting the behavior of *individual* participants.

The use of a construct as an explanatory tool thus differs from its use as a predictive tool. But the ability of a theoretical construct to explain observations should arguably count as much as its ability to predict them. Indeed, a construct simply cannot be supported evidentially other than by explaining *and* predicting observations (Klein et al., 2018). This means the proportion of erroneous predictions (γ , δ) should *match* the statistical error-rates (α , β). This insight is basic to the Hempel-Oppenheim schema (Hempel, 1970), where explanation and prediction are recognized as equally important.

In behavioral science, however, theoretical constructs are typically evaluated asymmetrically, placing the focus on *explanation*. But the explanation error-rates (α , β) can be reduced by simply increasing the statistical parameter N ; whereas the only way to reduce the prediction error-rates (γ , δ) is to increase the theoretical parameter. The focus on explanation, moreover, goes along with an *inductive* strategy. Here, the goal is to explain what was already observed. A finite number of (statistically significant) observations, however, cannot generate a predictively adequate construct automatically. As a method, induction can serve to *compare* theoretical constructs. But induction cannot construct them. (This is known as *Hume's problem of induction*; Rickles, 2020).

With a theoretical construct's ability to adequately predict observations being undervalued, behavioral science appears to undervalue *deductive* reasoning. Physicists, by contrast, pursue a deductive strategy to theory construction and evaluation. Here, the ability of a theoretical construct to predict is as important as its ability to explain. This strategy is precisely what necessitates that the SEOI for a theoretically worthwhile parameter be defined. Indeed, the importance of the deductive strategy for theory construction cannot be overstated: "[...] not specifying a SEOI [smallest effect of interest] for research questions at all will severely hinder theoretical progress" (Lakens, Scheel & Isager, 2018, 267).

A deductive strategy recommends predicting a *large* effect because this makes small samples sufficient to probabilistically verify that the predicted effect is stable (if it is). If one were to predict the *huge* effect $d=3.3$, for instance, then a t -test under $\alpha=\beta=0.05$ would merely require $N_{\text{MIN}}=N_1=N_2=4$, or the same number for $d=4.4$ under $\alpha=\beta=0.01$. Thus, if the proportion of erroneous predictions (γ , δ) shall match the statistical error-rates (α , β)—making explanation and prediction equally important—a theoretical construct should generally predict a *very large* effect.

But this starkly contrasts with the *small to medium* effects that behavioral science journals tend to report. The typical effect size-range of $0.20 \leq d \leq 0.60$ (Bakker, et al., 2012) indeed suggests a preference for a *vague* explanatory construct, as opposed to one that predicts new observations adequately. Meehl (1990) referred to the difference between the *modus operandi* in physics and behavioral science as a methodological paradox. That paradox would seem to persist (even if a properly planned study point-specifies d theoretically and then deduces N_{MIN}) until behavioral science—rather than discovering significant deviations from random—develops empirically adequate point-predictive theoretical constructs (Witte & Zenker, 2017). Discoveries, of course, help to develop a “feeling” for the effect. But the evaluation of a predicted effect requires a theoretical construct.

4.2.5 Upshot

NPTT’s α - and β -error-rates specify the probability of erroneously inferring that a hypothetical effect is true (Hubbard & Bayarri, 2008). Because only the explanation error-rates vary with N , a theoretical construct that states a hypothetically true effect incurs prediction error-rates that exceed the explanation error-rates. The scientific game against nature, however, is played fairly, only if the error-rates of theory-based applications and interventions (γ , δ) match the proportion of misleading data if a true effect is explained (α , β). To evaluate a theoretical construct, therefore, explanation and prediction are equally important.

Cohen’s d -measure, however, exclusively tracks the explanation error-rates, which are embedded into a significance test. But the d -measure ignores the prediction error-rates that evaluate the combined area *not* shared by two population distributions. This makes Cohen’s effect size classification (*small*, *medium*, *large* effect) inadequate to evaluate the scientific relevance of theoretically predicted effects. Particularly a construct that states a true explanation but fails to predict new data, is a *weak* construct that merely offers a glance at the SEOI. Weak constructs, however, are rather normal in behavioral science. Here, the ability to theoretically organize reality is poor because extant theoretical constructs are generally too weak to model *individual* behavior (single observations). They rather model *average* behavior given random fluctuations.

Presumably, stronger theories (that state larger effects) are unavailable because the vast heterogeneity of observations around a mean suggests that theoretically non-modelled influences cause individual behavior to deviate from the predicted mean. These deviations invariably collect as random fluctuations around the mean. The mean thus becomes an “error absorbing” parameter (*central limit theorem*). Because random fluctuations are always at play, a theoretical construct should be evaluated by treating *only* the mean as a theoretically modelled parameter (rather than the mean plus the variance of behavioral responses). The more informative this construct is, the better can observations be explained, and the smaller are the expected prediction error-rates.

All along, of course, statistically estimating an observed effect is conceptually distinct from predictively applying a corroborated theoretical construct. Just as explaining or predicting *individual* behavioral responses is distinct from explaining or predicting the *average* behavioral response.⁴

5. The practical aspect

All of this suggests that treatments or interventions with *large* effects are preferable, because large effects have comparatively greater utility relative to the status quo. By contrast, arguing in favor of *small* effects as practically relevant effects presupposes not only the effect’s utility but also its stability. For example, Poole-Wilson et al. (2003) report a 29% death rate for patients who are treated with cardiovascular drug A *vs.* a 35% death rate for drug B. An effect size difference of 6% is less than a small effect. And the utility difference between life and death is arguably *infinite*, trumping what is measured on a finite scale. What specifically disqualifies the effect size difference as the right evaluative criterion thus is the negative utility of death. Rather, if the effect of

⁴ An example is a “quick-test” for a viral infection. The test’s explanation error-rates can be determined by comparing a *sample* of quick-tests results against those of a more sophisticated test (e.g., a PCR test). The theoretical significance of developing a quick-test already stands or falls with the ability to outperform a chance result *on average*. Medical contexts nevertheless require a virus-test that predicts an infection correctly in each single case. This is achieved by measuring the *individual* viral load.

each drug is non-zero, one would ascribe a preference for drug A over drug B to the infinite expected utility of reducing the death rate.

This preference can be modeled as the expected value of the corresponding treatment. Here, ‘ P ’ is the probability of a positive outcome (“life”), ‘ U_L ’ is the outcome’s ascribed utility, ‘ $(1-P)$ ’ the probability of a negative outcome (“death”), and ‘ $-U_D$ ’ the ascribed loss, or negative utility. The expected value, E , for drugs the A and B then is:

$$E_A(U)=P_A \cdot U_L+(1-P_A) \cdot (-U_D), \text{ and } E_B(U)=P_B \cdot U_L+(1-P_B) \cdot (-U_D) \quad (1)$$

The death-rates for drugs A and B hence are $(1-P_A)=0.29 < (1-P_B)=0.35$, while $-U_D=\text{about}-\infty$ (“negative infinity”). According to (1), then, the dominating choice is drug A. The preference of drug A over drug B thus has a rational model, given that a non-zero effect size lowers the probability of death. But the smaller the death-rate difference, the more error-prone it is. (Eventually, the *true* death-rate difference may even lie in the opposite direction.) On the assumption that the death-rate difference is *stable* if the associated error-rate is $<1\%$ ($\alpha=\beta<0.01$), demonstrating its stability would hence require $N_{MIN1}=N_{MIN2}=6100$ experimental participants (Faul et al., 2007). Any smaller sample simply leaves the true death-rate difference too uncertain.

This shows that a practically relevant effect must be stable. Conversely, a stable non-zero effect can vary freely, only if choices depend *exclusively* on their ascribed utility. Under the practical aspect, then, an effect’s expected value is estimated as a function of the effect’s error-rates and its expected utility. This contrasts with the description of an effect as an inductive parameter, a description that depends on measurement contingencies or error-prone predictions. Since only a *stable* probability estimate will warrant attending to a very small effect, and as only a very large sample yields the CIs to statistically exclude the inverse effect, a very small stable effect may well be practically relevant. But the larger the effect is, the easier it is to (fallibly) know in evidence-based ways that it *is* practically relevant.

Using utility considerations in evaluative research-planning generally requires that four specific utilities be associated with a study-outcome (Miller & Ulrich, 2016):

utilities for *false positives* (erroneously rejecting H_0), *true positives* (correctly accepting H_1), *false negatives* (erroneously rejecting H_1), and *true negatives* (correctly accepting H_0). These four additional parameters, however, are potentially confusing rather than helpful, because one must avoid that the utilities *alone* justify the evaluation. Implicit or explicit use of utilities nevertheless goes a long way towards explaining the tolerance for low test-power in behavioral science. Suitably specified, after all, utilities can make it *seem* as if a research-output is maximized, even though the discipline is not improved theoretically (Miller & Ulrich, 2016, 671; Table 2).

A case in point is the small effect of the IAT (see Sect. 1). The IAT is defensible *only* on the utility-based claim that its effect applies in a very large sample. Given “big data,” of course, a very small effect can quickly become useful for many individuals (e.g., Matz et al., 2017). This holds whether utilities are based on severe consequences for a single individual (e.g., personal health context) or mild consequences for many individuals (e.g., public policy context). In cases similar to the IAT, however, absence of a theoretical explanation implies knowing *that* the effect occurs, but not *why*.

If each mistaken decision has the same disutility, and if each correct decision has an inverse positive utility, then the most convincing utilities probably are the error-rates themselves. At any rate, a decision-theoretic treatment does not necessitate assumptions besides those underlying a statistical theory. So, if the goal is to deduce empirically adequate predictions—and it should be—then not only would a theoretical construct that states a comparatively large effect be comparatively more useful. The construct’s empirical adequacy could also be assessed with a comparatively smaller sample (Witte & Zenker, 2017; Krefeld-Schwalb, Witte & Zenker, 2018). To plan a construct-corroborating experiment, by contrast, utilities are of no use at all.

6. Conclusion

A transparent evaluation of an effect should distinguish at least three aspects: the size of effect (*statistical aspect*), how well a theory models it (*theoretical aspect*), and its utility (*practical aspect*). Under the statistical aspect, one does not really evaluate but rather describes the effect as a parameter estimate. Under the theoretical aspect, one evaluates the empirical adequacy of a theoretical construct by assessing whether a prediction fits with *new* observations. Under the practical aspect, finally, one evaluates

the effect's utility. Keeping these three aspects apart clarifies what the one-dimensional evaluation (as a *small*, *medium*, or *large* effect) suggested by Cohen's *d*-measure oversimplifies. Although we discussed only simple hypotheses (that state a mean difference) the same holds for more complex hypotheses (e.g., correlations, multiple means, interactions, dimensional hypotheses, etc.). Moreover, an evaluation based on these three aspects remains distinct from a theory-test because it cites probabilities rather than likelihoods.

Two error-pairs quantify the uncertainty that is incurred when a theoretical construct *explains* observations (α -, β -error-rates) as opposed to when it *predicts* them (γ -, δ -error-rates). But the uncertainty varies with *N* only in case of explanations. The 'smallest effect of interest' (SEOI) for an empirically adequate prediction is therefore *larger* than the SEOI for an empirically adequate explanation. Consequently, a small effect generally lacks *theoretical relevance*, even though expected utility considerations can warrant its *practical relevance*. Given that the homogeneous share of observations in behavioral science comprises *small* effects, defending these as relevant thus requires invoking the practical aspect. That said, a practical decision that is based on what a typical theoretical construct predicts for *individual* behavioral responses presumably outperforms a fair coin-toss but slightly.

Acknowledgements/Declarations

E.W. and F.Z. declare that they have no conflict of interest with respect to this publication. E.W. wrote a first draft that FZ and E.W. jointly developed, and which F.Z. edited. Both authors approved the final submitted version.

References

- Anderson, N.H. (1991a). *Contributions to information integration theory*. Vol. I: *Cognition*. Hillsdale: LEA.
- Anderson, N.H. (1991b). *Contributions to information integration theory*. Vol. II: *Social*. Hillsdale: LEA.
- Anderson, N.H. (1991c). *Contributions to information integration theory*. Vol. III: *Developmental*. Hillsdale: LEA.
- Anderson, N.H. (1996). *A functional theory of cognition*. Mahwah: LEA.

- Bamber, D., & van Santen, J.P.H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, 44, 20-40.
- Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, 100, 603-617.
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-437.
- Bakker, M., Dijk, van A., & Wicherts, J.M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543-554.
- Borsboom, D., & Mellenbergh, G.J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence*, 30, 505-514.
- Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological review*, 111(4), 1061-1071.
- Boyce, C.J., Daly, M., Hounkpatin, H.O. & Wood, A.M. (2017). Money may buy happiness, but often so little that it doesn't matter. *Psychological Science*, 28, 544-546.
- Cafri, G., Kromrey, J.D., & Brannick, M.T. (2010). A Meta-Meta-Analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research*, 45, 239-270.
- Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of abnormal and social psychology*, 65, 145-153.
- Cohen, J. (1977, rev. ed). *Statistical power analysis for the behavioral sciences*. London: Academic Press.
- Duhem, P. (2015) [1908]. *To save the phenomena: An essay on the idea of physical theory from Plato to Galileo*. Chicago, ILL.: University of Chicago Press.
- Eckes, Th., & Six, B. (1994). Fakten und Fiktionen in der Einstellungs-Verhaltensforschung: Eine Meta-Analyse. [Fact and fiction in attitude-behavior research: A meta-analysis.] *Zeitschrift für Sozialpsychologie*, 25(4), 253-271.
- 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.

- Ferguson, C.J. (2009). An Effect Size Primer: A Guide for Clinicians and Researchers. *Professional Psychology: Research and Practice*, 40(5), 532-538.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston: Row, Patterson, and Company.
- Fisher, R.A. (1956). *Statistical methods and scientific inference*. New York: Hafner.
- Fildes, R., Nikolopoulos, K., Crone, S.F., & Syntetos, A.A. (2008). Forecasting and Operational Research: A Review. *Journal of the Operational Research Society*, 59, 1150-1172.
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456-465.
- Forster, M.R. (2000). Key concepts in model selection: Performance and Generalizability. *Journal of Mathematical Psychology*, 44, 205-231.
- Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106, 746-762.
- Greenwald, A.G., Banaji, M.R., & Nosek, B.A. (2015). Statistical small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, 108(4), 553-561.
- Grissom, R.J., & Kim, J.J. (2012²). *Effect sizes for research*. New York: Routledge.
- Hempel, C. G. (1970). On the “standard conception” of scientific theories. In: M. Radner & S. Winokur (eds.), *Analysis of theories and methods in physics and psychology* (pp. 142-163). Minnesota: Minnesota Studies in Philosophy of Science, IV.
- Hubbard, R. & Bayarri, M. J. (2008). Confusion of measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *American Statistician*, 57(3), 171-178.
- Huberty, C.J. (2002). A history of effect size indices. *Educational and psychological measurement*, 62, 227-240.
- 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. <https://doi.org/10.17605/OSF.IO/8CD4R>

- Krefeld-Schwalb, A; Witte, E.H., and Zenker, F. (2018). Hypothesis-testing demands trustworthy data—a simulation approach to statistical inference advocating the research program strategy. *Frontiers in Psychology*, 9, 460.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, article 863, doi: 10.3389/fpsyg.2013.00863.
- 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. (2021). Heterogeneity of research results: A new Perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, 16(2), 358–376.
- Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
- McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, 11(4), 386-401. <https://doi.org/10.1037/1082-989X.114.386>
- Matz, S.C., Gladstone, J.J., & Stillwell, D. (2017). In a world of big data, small effects can still matter: A reply to Boyce, Daly, Hounkpatin, and Wood (2017). *Psychological Science*, 28(4), 547-550.
- Meehl, P.E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1, 108-141 and 173-180.
- Meehl, P.E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numeral predictions. In: L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.). *What if there were no significance tests?* (pp. 393- 425). Mahwah: Erlbaum.
- Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, 7, 109-117.
- Miller, J. & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, 11, 664-691. <https://doi.org/10.1177/1745691616649170>

- Myung, I.J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44, 190-204.
- Neyman, J., & Pearson, E.S. (1967). *Joint statistical papers*. Cambridge: Cambridge University Press.
- Olsson-Collentine, A., Wicherts, J.M. & van Assen, M.A.L.M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, 146(10), 922–940. <https://doi.org/10.1037/bul0000294>
- Oswald, F.L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P.E. (2013). Predicting ethnic and racial discrimination: A meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105, 171-192.
- Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242.
- Poole-Wilson P.A., Swedberg, K., Cleland, J.G., Di Lenarda, A., Hanrath, P., Komajda, M., Lubsen, J., Lutiger, B., Metra, M., Remme, W.J., & Torp-Pedersen, C. (2003). Comparison of Carvedilol and metoprolol on clinical outcomes in patients with chronic heart failure in the Carvedilol or Metoprolol European Trial (COMET): randomized controlled trial. *Lancet*, 362, 7-13.
- 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.
- Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, 45, 775-777.
- Rosenthal, R., & Rubin, D.B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, 86, 1165-1168.
- Rosenthal, R. & Rubin, D. B. (1982). A simple general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166-169.
- Schäfer, Th. (2018, November 23). Effect sizes in psychological research. Retrieved from osf.io/t8uvc.
- Schauer, J.M. & Hedges, L.V. (2020, April, 9). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, 146(8), 701–719. <https://doi.apa.org/doi/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>
- Schmidt, F. L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262-274.
- Schmidt, F.L., Oh, I., & Hayes, T.L. (2009). Fixed versus random-effect models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, 62, 97-128.
- Schulze, R. (2004). *Meta-Analysis. A comparison of approaches*. Göttingen: Hogrefe & Huber.
- Schwartz, S.H., Verkasalo, M., Antonovsky, A., & Sagiv, L. (1997). Value priorities and social desirability: Much substance, some style. *British Journal of Social Psychology*, 36, 3-18.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, 22(11), 1359-1366.
- Sixtl, F. (1993). *Der Mythos des Mittelwertes. [The myth of the mean.]* München: Oldenbourg.
- 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>
- Viechtbauer, W. (2007). Accounting for Heterogeneity via Random-Effects Models and Moderator Analyses in Meta-Analysis. *Zeitschrift für Psychologie / Journal of Psychology*, 215, 104-121. <https://doi.org/10.1027/0044-3409.215.2.104>
- Wilkinson, L., & APA Task Force (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
- Witte, E.H. (1979). *Das Verhalten in Gruppensituationen. Ein theoretisches Konzept*. Göttingen: Hogrefe.
- Witte, E. H. (1980). Die Theorie der kognitiven Dissonanz und das Konzept der Informationsintegration: Zur Entwicklung einer gemeinsamen Theorie. In E.H. Witte (Hrsg.), *Beiträge zur Sozialpsychologie* (S. 233–251). Weinheim: Beltz.

- Witte, E. H. (1994). Minority influences and innovations: The search for integrated explanation of psychological and sociological models. In S. Moscovici., A. Mucchi-Faina & A. Maass (Eds.), *Minority influence* (pp. 67–93). Chicago: Nelson-Hall.
- Witte, E.H. (1996). The extended group situation theory (EGST): Explaining the amount of change. In: E.H. Witte & J.H. Davis (Eds.). *Understanding group behavior* (Vol. 1) (pp. 253-291). Mahwah: Erlbaum.
- Witte, E.H., & Heitkamp, I. (2006). Quantitative Rekonstruktionen (Retrognosen) als Instrument der Theorienbildung und Theorienprüfung in der Sozialpsychologie. *Zeitschrift für Sozialpsychologie*, 37 (3), 205–214.
- Witte, E. H. & Zenker, F. (2016) Reconstructing recent work on macrosocial stress as a research program. *Basic and Applied Social Psychology*, 38(6), 301-307. DOI: 10.1080/01973533.2016.1207077
- Witte, E.H., & Zenker, F. (2017). From discovery to justification: Outline of an ideal research program in empirical psychology. *Frontiers in Psychology*, 8:1847. <https://doi.org/10.3389/fpsyg.2017.01847>
- Wulff, D.U., & van den Bos, W. (2017). Modeling Choices in Delay Discounting. *Psychological Science*, 1-5. <https://doi.org/10.1177/2F0956797616664342>
- Young, F.W. (1972). A model for polynomial conjoint analysis algorithms. In: R.N. Sheppard, A.K. Romney; & S.B. Nerlove (eds.). *Multidimensional Scaling. Theory and applications in the behavioral sciences*. Vol. I (69-104). New York: Seminar Press.
- Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, 44, 41-61.

Word count: 9200