EJOP Europe's Journal of Psychology PsychOpen GOLD

Theoretical Contributions

Check for updates

Statistical Tests Should Inform Behaviour in Theory and Practice

Michael Höfler [1], Sebastian Trautmann [2], Philipp Kanske [1,3], Robert Miller [4]

[1] *Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, TUD Dresden University of Technology, Dresden, Germany.* **[2]** *Institute for Clinical Psychology and Psychotherapy, Department of Psychology, Medical School Hamburg, Hamburg, Germany.* **[3]** *Department of Psychology, Faculty of Psychology and Educational Sciences, Babeş-Bolyai University, Cluj-Napoca, Romania.* **[4]** *Chair of Psychological Methods, Psychologische Hochschule Berlin, Berlin, Germany.*

Europe's Journal of Psychology, 2026, Vol. 22(4), 621–646, https://doi.org/10.5964/ejop.18963

Received: 2025-07-21 • **Accepted:** 2026-04-20 • **Published (VoR):** 2026-11-23

Handling Editor: Johannes Karl, Stanford University, Stanford, USA

Corresponding Author: Michael Höfler, Institute of Clinical Psychology and Psychotherapy, TUD Dresden University of Technology, Chemnitzer Straße 46, 01187 Dresden, Germany. Phone: +49-351-469321. E-mail: michael.hoefler@tu-dresden.de

Abstract

Objectives: Statistical hypothesis testing cannot establish a hypothesis as certainly—or even approximately—true, but it is valuable in guiding behaviour *as if* it were true. To make decisions in theory and practice, statistical testing in psychology should weigh the Type I (α) and Type II (β) error probabilities according to their respective *costs*.

Methods: We outline costs and how to approach the α-β balance for six common decision-making questions: (1) Should a hypothesis be considered tentatively accepted? (2) Should a model or theory cover a newly proposed effect? (3) Should a new hypothesis be proposed? (4) Should a new intervention be introduced? (5) Are the psychometric properties of an instrument sufficient to use it? (6) Are the assumptions of a statistical method met so that it may be used? We give a worked example with real cost figures for question (4).

Results: We argue that parsimony—prioritizing low α over low β—often serves as a useful default stance in questions (1) to (5). This resembles the general preference for focusing on fewer, larger and more robust effects to allocate resources efficiently and to make scientific communication more reliable.

Conclusions: The parsimony default invites transparent counterarguments to override it, justifying a preference for small β. Examples include adding promising intervention components with little risk of adverse effects to well-established interventions. As with (6), statistical tests can

This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License, CC BY 4.0, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

sometimes introduce unnecessary errors when a sound decision can be made on substantive grounds alone, for example when setting up a model to test a given hypothesis.

Keywords

Statistical Hypothesis Testing, Type I Error (α), Type II Error (β), Error Balancing, Statistical Decision-Making, Applied Statistics in Psychology

In psychology, social, health, and other life sciences, there is an enduring debate about how to conduct statistical inference and whether the common α error threshold of 0.05 should be abandoned (Benjamin et al., 2018; Cohen, 1994; Lakens et al., 2018a; Lykken, 1968; Mayo, 2018; McShane et al., 2019; Wasserstein & Lazar, 2016). Several authors have argued that it is overly ambitious to assume that statistical hypothesis testing—regardless of whether frequentist or Bayesian methods are used—can determine with certainty whether a hypothesis is true or false (Cartwright et al., 2022; Gigerenzer, 2004; Greenland, 2019; Meehl, 1990; Schneider, 2015). The main argument is that a complex reality—such as that studied in psychology—can never be accurately represented, even approximately, by statistical models, and thus cannot be probed with a single test. In the philosophy of science, this insight is long known as the Duhem–Quine thesis (Duhem, 1954; Quine, 1951). It implies that any hypothesis test based on a statistical model relies on auxiliary assumptions, such as the absence of any bias (Cartwright et al., 2022; Greenland, 2005; Meehl, 1990). 'True' in practical terms may only claim that a hypothesis is at least approximately true or merely useful—for example, the statement 'the therapy is effective' may not literally apply to every individual, but it might on average. Gigerenzer and colleagues (2004) concluded: 'Accepting a hypothesis does not imply that you believe in it; it only means that you act as if it were true.' Thus, statistical inference should *inform decisions about behaviour* in theory or practice (Aisbett, 2024; Dawid, 2021; Greenland, 2021; Killeen, 2006; Lakens et al., 2018a; Maier & Lakens, 2022; Petersen; 2024; Rafi & Greenland, 2020). Aiming for the best decision involves not only the two error probabilities—Type I (α), the risk of wrongly supporting a hypothesis, and Type II (β), the risk of wrongly rejecting one—but also the associated costs of both errors. According to statistical decision theory (Wald, 1950), the decision with the lower expected cost is preferred.

This article posits that psychologists implicitly employ statistical tests in their decision-making processes. It also assumes that greater awareness of the underlying decisional questions, along with at least a basic understanding of the costs associated with the two possible errors of a test, will lead to better-informed decisions. Cost considerations are expected to vary across multiple levels: the research question, the broader field, and the specific context. We examine six decisional questions that we consider particularly relevant and sufficiently uniform to allow researchers to outline the types of costs involved and to consider how they might be weighted:

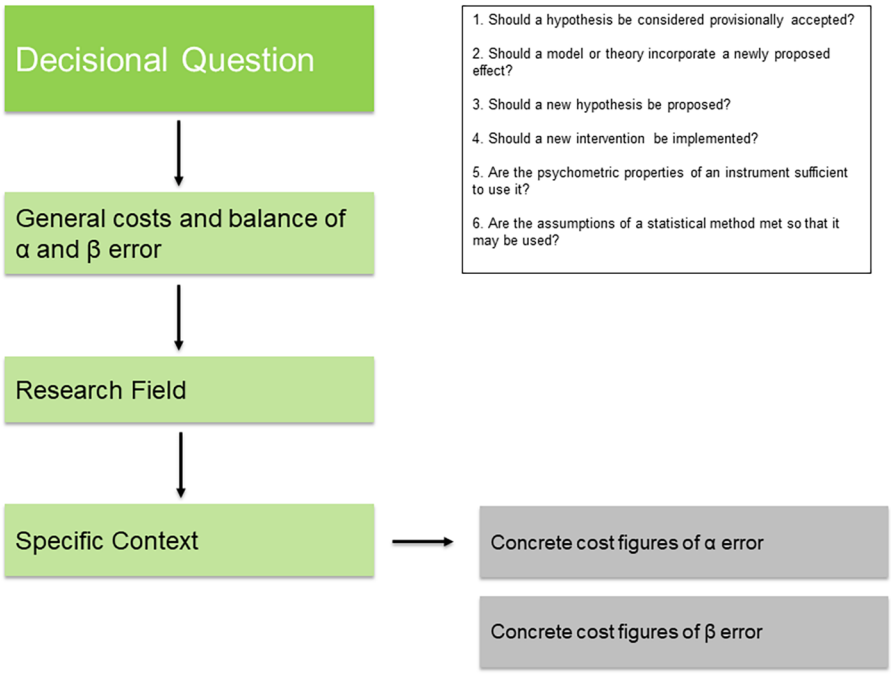
PsychOpen GOLD

1. Should a Hypothesis be Considered Provisionally Accepted?
2. Should a Model or Theory Incorporate a Newly Proposed Effect?
3. Should a New Hypothesis be Proposed?
4. Should a New Intervention be Implemented?
5. Are the Psychometric Properties of an Instrument Sufficient to Use it?
6. Are the Assumptions of a Statistical Method Met so that it may be Used?

Figure 1 summarizes our approach in a conceptual scheme:

Figure 1

A Conceptual Scheme of Addressing Error Costs Across Multiple Levels of Abstraction



In this article, we focus mainly on the first level. However, to illustrate how to approach the α–β balance in practice, we provide a worked example for Question 4 with concrete cost figures. After that, we mention three overarching aspects of statistical testing that apply across all six questions and outline how cost considerations can guide behaviour in these cases: making decisions on different questions using the same test, making decisions in the presence of multiple testing, and determining whether a new study is required when the current study appears inconclusive. We will conclude by discussing decisions that should not be based on statistical tests at all—since doing so would only

PsychOpen GOLD

introduce unnecessary errors—and should instead be grounded solely in substantive considerations.

In line with the formulation of decision questions 1-5, we focus on *novel* as opposed to well-established findings which might later be contested. Thus, we refrain from addressing the extensive topic of whether to discard established hypotheses in light of new evidence.

We will often refer to the simple example of the null hypothesis of no effect and the alternative hypothesis of an effect. However, the arguments similarly apply to other kinds of statistical tests where α and β can be fixed before access to data: one-sided tests, point versus point tests (e.g. correlation = 0 against correlation = 0.5, Neyman & Pearson, 1933) and equivalence tests where the alternative hypothesis states the absence of an effect of relevant size (here, α and β must be swapped for our arguments to apply). (Goeman et al., 2010). Although the historical development of Fisherian significance testing (Fisher, 1925) and Neyman–Pearson testing (Neyman & Pearson, 1933) predates the formalization of decision theory (Wald, 1950) and involves distinct philosophical underpinnings, their conceptions are nevertheless relevant for decision-making. We argue that all six decision questions addressed here, regardless of the type of statistical tests used, have consequences that involve costs and therefore call for balancing α and β. We phrase these questions as binary alternatives that are both theoretically and practically relevant, reflecting the inherent binarity of statistical tests (Uygun Tunç et al., 2023).

We discuss the balance of α and β from the conceptual standpoint of study and analysis design. This means that, before collecting data, their values should be preregistered to hold researchers accountable for them (Lakens et al., 2018a; Lakens et al., 2024; Mayo, 2018). While α can be freely chosen in a statistical test (as it pertains to planning the analysis and its decision rule on the hypothesis, e.g., $p < \alpha$), β is strongly influenced by sample size, which must be carefully planned to achieve the desired power, 1 - β. A power determination aimed at detecting effects in terms of the smallest effect size of interest (SESOI) aligns with the goal of identifying effects of theoretical or practical relevance with the desired probability, 1 – β (Altman & Bland, 1995; Lakens et al., 2018b; Murphy & Myors, 1999; Neyman & Pearson, 1933). This aims to avoid inconclusive results (Goeman et al., 2010) from the outset; that is, results that do neither support the alternative hypothesis (e.g. existence of an effect) nor the null hypothesis (e.g. absence of an effect).

Our main focus is on designing studies with strict adherence to α and β, but we occasionally touch on questionable research practices, such as p-hacking, that undermine this adherence.

PsychOpen GOLD

Statistical Decision Theory

Statistical decision theory was first proposed in 1950 (Wald, 1950), further developed (DeGroot, 1970; Raiffa & Schlaifer, 1961) and applied (Dawid, 2021). It assumes a set of possible actions, each with consequences occurring with certain probabilities (a probability distribution). A statistical test is a decision between two hypotheses—choosing which to prefer (e.g. acting as if there is no effect versus an effect). The consequences are the common errors: falsely assuming an effect or falsely rejecting one, with probabilities α and β. Following the Neyman-Pearson tradition, α and β must be fixed and independent of the data. This ensures rigor and falsifiability of both hypotheses, no matter what is observed (Mayo, 2018; Neyman & Pearson, 1933). However, the replication crisis has demonstrated that falsifiability is often undermined by questionable research practices, such as p-hacking and HARKing (Nagy et al., 2025). Statistical decision theory adds that the costs of these errors, C_α and C_β, must be incorporated into the choice of α and β. The costs must be known or, in practice, specified by informed guess to guide the choice and balance between them (Greenland, 2019).

The expected cost due to α equals $\alpha * C_\alpha$, and the expected cost due to β is $\beta * C_\beta$. Regarded as a decision problem, a statistical test is about minimizing the sum of these losses—the total expected cost (EC):

$$\min_{\alpha,\beta} EC = \alpha * C_\alpha + \beta * C_\beta.$$

Minimizing the expected cost is achieved when both Type I (α) and Type II (β) errors are kept small simultaneously. However, this is effortful in practice, as reducing α results in an increase in β for a given data set. In study design, this trade-off can be mitigated by increasing the sample size (Neyman & Pearson, 1933). Conversely, with a very small β achieved through a large sample size, a test may pass even a stringent α level—yet still detect trivially small effects (Maier & Lakens, 2022; Spanos, 2013). Thus, weighing α and β to minimize costs is a sound approach. The expected cost (EC) equation is instructive here, with the optimal β/α ratio given by:

$$\beta / \alpha = C_\alpha / C_\beta.$$

This relation implies:

- For equal costs, $C_\alpha = C_\beta$, one should choose $\alpha = \beta$.
- If $C_\beta > C_\alpha$, the test should tolerate higher α than β.
- If $C_\alpha > C_\beta$, the test should tolerate higher β than α.

For example, the common choice of $\alpha = .05$ and $\beta = 0.20$ implicitly reflects a decision-theoretic assumption that a Type II error (falsely assuming no effect) is four times less costly than a Type I error (falsely assuming an effect), under the assumption of equal prior probabilities of H_0 and H_1. However, these considerations are genuinely *frequentist*.

PsychOpen GOLD

Once the sample size is set, β depends strongly on it, and frequentist tests are designed with fixed α and β values. Strictly speaking, these values are not probabilities in the usual sense; rather, they describe the long-term error rates associated with H_0 and H_1 (Maier & Lakens, 2022, and references therein). We will see that deviations from these assumptions—particularly regarding equal prior probabilities— are highly relevant for decision questions 1-5 and can substantially affect the optimal balance between α and β. This naturally leads to the extended, Bayesian version of the expected cost formula (Berger, 1985):

$$EC = P(H_0) * \alpha * C_\alpha + P(H_1) * \beta * C_\beta.$$

The optimal error ratio becomes:

$$\beta / \alpha = (P(H_0) * C_\alpha) / (P(H_1) * C_\beta).$$

Here, the prior probabilities $P(H_0)$ and $P(H_1)$ weigh the cost terms. For instance, if an effect (H_1) is considered a priori unlikely—say, $P(H_1) = 0.1$ (and $P(H_0) = 0.9$)—and the costs of both error types are equal, then α should be set to 1/9 of β to reflect this scepticism. A lower α imposes a severe threshold for establishing such an effect, while a β nine times as high may be tolerated for missing one. Crucially, weighting of error costs by the prior plausibility of the hypotheses should be done even in common frequentist testing (Maier & Lakens, 2022).

Bayesian approaches, by contrast, focus on the probability of a hypothesis being true (given the model and data), rather than on fixed long-run error rates. Conceptually, this contrasts with frequentist methods, where α and β represent fixed probabilities *conditioned* on the truth or falsity of the null hypothesis. Technically, in Bayesian inference, the probability of being wrong is data-dependent, as evidence is incorporated through the likelihood and prior distributions. Since methods such as the Bayes factor accumulate evidence sequentially, their results may depend on the order in which data arrive or on whether data collection is stopped early (for instance, when the Bayes factor exceeds a common threshold such as > 10 or < 1/10). These dependencies arise because Bayesian updating treats each observation as incrementally modifying the likelihood and posterior odds for the hypotheses, so different data streams or stopping rules can yield different evidential conclusions. For example, a Bayes factor > 10 in favour of an effect (H_1) may occur early in data collection, helping to keep α low, while β might be low because stopping at the first strong piece of evidence in favour of an effect does not imply that the test procedure would have detected a real effect reliably. Certain decision-theoretic Bayesian frameworks avoid the dependence on interim looks, stopping rules, or data order by specifying a loss function in advance and basing decisions solely on the final posterior distribution (Berger, 1985; Gelfand & Wang, 2002; Lindley, 1985; Mayo, 2018; Robert, 2007). Strictly speaking, for both Bayesian and frequentist methods, fixing β also

PsychOpen GOLD

requires accounting for its dependence on α, and optimally balancing the two under cost considerations is inherently complex and iterative (Maier & Lakens, 2022).

Instead of delving into technical details or providing numerical examples, we outline how—whether within traditional frequentist testing or these specific Bayesian approaches—basic assumptions about the relative costs of α and β can provide a principled foundation for decision-making. This should replace the common practice of seemingly circumventing such choices—that is, making them by convention and without justification.

Statistical Tests, α and β, and the Decisions They Inform

The wide range of purposes for which statistical tests are used and the need to balance α and β according to their respective costs have long been noted (Aisbett, 2024; Gigerenzer, 2018; Greenland, 2022; Maier & Lakens, 2022; Meehl, 1978; Pearson, 1962). In the following, we will discuss this separately for the six decision questions listed in the introduction.

1. Should a Hypothesis be Considered Provisionally Accepted? (Confirmatory Research)

This decision seems to underlie many uses of statistical tests, where no other purpose can be inferred from the reported background of an analysis (Benjamin et al., 2018; Ioannidis, 2005; Lindley, 1985; Maier & Lakens, 2022). Making the costs of wrong decisions transparent becomes especially important when one must behave upon provisional acceptance or provisional nonacceptance. In confirmatory research and the tradition of Popperian falsification (Popper, 1959; Mayo, 2018) it serves the goal of rigidly testing a new hypothesis —such as the existence of an effect. Passing such testing justifies further investigation until new, more convincing evidence might suggest otherwise (Höfler et al., 2022). Provisional acceptance may warrant subsequent research into how, why, or for whom the effect occurs—for example, by exploring mechanisms, moderators, mediators, or boundary conditions. Other studies might again examine the possibility that the effect does not actually exist—for instance, by using a more accurate experimental paradigm. Confirmatory research should take into account scientific communication and the costs of establishing a non-existent effect in the literature versus not establishing an existing effect. In the first case, new research is based on a false claim and time, financial and human resources (of study participants and researchers) are misdirected. Costs also involve the need of a perhaps expensive replication study in case of doubts on a falsely established hypothesis (Maier & Lakens, 2022). If a true hypothesis is missed, resources are also wasted because studies that could build on it are delayed, instead effort is put

PsychOpen GOLD

into further studies before establishing it. Lindley (1985). Ioannidis (2005) and Maier and Lakens (2022) suggest a smaller α (e.g. 0.01 instead of 0.05) when the prior probability of an effect is low. However, this comes at the expense of a higher β—or requires costly large sample sizes to keep β low (Benjamin et al., 2018; Lakens, 2022). We agree with others that much psychological research would benefit from focusing on fewer and larger effects that are more robust, predictive (thus falsifiable) and replicable (Fraley & Vazire, 2014; Meehl, 1967). Increasing β would primarily come with the often acceptable cost of overlooking more small effects (Lakens et al., 2018a; McShane et al., 2019). In contrast, small genetic effects—or analogous psychological effects—have been defended on the grounds that they can accumulate across the genome or over time (Gordon, 2021; Götz et al., 2022). However, this argument commonly refers to the *aggregate* of such effects, and the justification for them rests on the predictive and explanatory utility of the combined influence. One example of a potentially meaningful aggregate is the genetic predisposition to the Big Five personality traits: genome-wide association studies suggest that several genes contribute to these predispositions, though collectively they explain only a small portion of the variance in the traits (Schwaba et al., 2025).

Benjamin et al. (2018, criticized by Lakens et al., 2018a) recommended a stricter α of 0.005 (based on correspondence with Bayes factors), while tolerating a β of at most 0.1. Independent of this reasoning, many researchers also argue that stricter α levels can help mitigate issues related to undisclosed analytical flexibility in practice (Ioannidis, 2005). We believe that this issue should instead be addressed by requiring transparency practices such as preregistration for a result to be accepted as confirming (Höfler et al., 2022). This would facilitate the α–β discussion by basing it on the actual α.

2. Should a Model or Theory Incorporate a Newly Proposed Effect? (Confirmatory Research)

If the hypothesis in 1. concerns the existence of an effect, this question adds a piece of context, namely the model or theory in which the effect would be embedded. The associated behavioural decision is whether or not to add the effect. Falsely including a newly proposed effect in a model or theory has several costs: reduced replicability due to overfitting the data (Yarkoni & Westfall, 2017), unnecessary theoretical complexity (Roberts & Pashler, 2000), and distraction from strong effects by treating spurious effects equally (MacCallum et al., 1992). Conversely, incorrectly omitting an effect reduces explanatory power—particularly in accounting for individual differences. Overly simplistic models have been criticized for generating fewer testable predictions, potentially hindering theoretical progress (Meehl, 1990).

One may agree that it is acceptable to omit weak effects from models (Vandekerckhove et al., 2018)—in statistical terms, to accept some bias in order to keep variance small (Gigerenzer et al., 2004)—and that the value of models lies in their explanatory and predictive capacity with just a few variables (Box, 1976). Then the risk of false

PsychOpen GOLD

inclusion outweighs the risk of false exclusion and similar recommendations as in 1 seem warranted as a default suggestion: Choose a small α and tolerate a larger β. A different choice might nevertheless be justified, if compelling arguments are stated in favour of small effects.

3. Should a New Hypothesis be Proposed? (Exploratory Research)

In research practice, this decisional question of exploratory research can hardly be distinguished from the question of testing a given hypothesis (confirmatory research, 1.). Content analyses show that similar wording is commonly used for both, and that statistical tests are applied not only to decide on a given hypothesis, but also to determine whether a new hypothesis should be proposed (Höfler et al., 2022, and references therein). This blending has contributed to the replication crisis and should be addressed again by requiring detailed and verified preregistration for an analysis to be accepted as confirmatory (Höfler et al., 2022).

In fact, exploratory data analysis concerns deciding whether a new hypothesis appears promising based on patterns found in the data. Statistical tests here serve a different, often only implicit goal: filtering the many patterns that may arise so that, ideally, only true, non-random, and substantively interesting patterns remain (Höfler et al., 2023). In research practice, confirmatory and exploratory questions are not only blended, but striking data patterns are often interpreted substantively (usually in confirmatory terms) even when they are not statistically significant at the conventional α level of 0.05—especially when they are labeled "trend significant" based on a post hoc threshold of 0.10. However, in such cases, they do not require a substantive explanation, as they can be accounted for by the most parsimonious explanation in science: chance. Thus, new findings should pass common thresholds for chance findings before a new hypothesis is claimed, thereby tolerating a larger β. They should often even withstand a more rigorous α, the more intense exploration has been used to find them and the less plausible they seem (Aisbett, 2024; Ioannidis, 2005; Maier & Lakens, 2022). While initial exploration should have the freedom to crawl through a dataset in search of every potentially interesting pattern, the abundance of such findings should be distilled to those most relevant and striking—guided by a theoretical framework that justifies such judgment. With only a few hypotheses left, these are more likely to pass a rigorous confirmatory test in a new data set. This serves scientific communication by focusing on just a few promising new hypotheses, and again draws attention to larger effects (Höfler et al., 2023).

However, the tentative preference for a lower α can be overcome in favour of a lower β when a new finding is both highly plausible and/or, if true, offers substantial potential benefits to individuals or the community with minimal associated risks. In such cases, adopting a more lenient α threshold to invite further confirmatory research

PsychOpen GOLD

can be justified. For instance, little initial evidence seems necessary to hypothesize that pre-booked appointments increase vaccination rates (Bonander et al., 2022).

4. Should a New Intervention be Implemented? (Confirmatory Research)

This question concerns the decision to behave as if an intervention is effective or not. Choosing to behave as if an intervention works when it does not (α error), such as a preventive measure or treatment, leads to unnecessary expenditure of financial and human resources, the mistaken attribution of adverse effects to the intervention, and the potential loss of opportunities to use genuinely effective interventions. Conversely, failing to implement an effective intervention or overlooking a genuine difference between the effects of various treatments (β error) delays essential public health actions. Such delay can lead to widespread health repercussions (e.g. a growing pandemic) and deny individuals access to beneficial therapies (Greenland, 2021), thereby perpetuating the status quo in research and clinical practice (Sugarman, 2016). These considerations may lead to very different recommendations depending on the type and magnitude of costs associated with an intervention. The complexity increases further when SESOI in terms of practical or clinical significance is incorporated into the question. Assessing clinical significance requires linking statistical effects to external benchmarks for intervention success, such as patient-reported improvements (Kirk, 1996). Likewise, decisions about whether to implement an intervention require evaluating a range of outcomes that may improve or worsen, encompassing short- and long-term effects and any potential adverse effects. We limit ourselves to selected examples from clinical psychology in which prioritizing a low β over a low α is warranted—serving as counterexamples to the parsimony principle outlined in Sections 1–3:

- **Testing a Harmless Supplement as an Addition to an Already Effective Therapy:** The effect of an established, evidence-based treatment might be enhanced by adding another component. If the added element is inexpensive and very likely harmless, using it even without strong prior evidence poses relatively little risk while offering the possibility of meaningful benefit. In medicine, this idea has been discussed in the context of supplementing a standard prophylactic antibiotic with a cheap, safe vitamin (Schulz & Grimes, 2005). A similar rationale could apply to promising adjunctive components to common psychotherapy protocols — for example, adding a brief course of motivational interviewing to cognitive behavioural therapy (Westra et al., 2009).
- **Antidepressants vs. Psychotherapies:** Real differences between the efficacy of antidepressants and psychotherapies may be overlooked, delaying their optimal integration into clinical practice (Furukawa et al., 2016; Sugarman, 2016). Replacing antidepressant therapy with psychotherapy may reduce costs in the long run with

PsychOpen GOLD

regard to quality of life, especially if psychotherapy provides more sustained benefits over time and requires less frequent intervention as patients improve (Ross et al., 2019).

- **Hormone Replacement Therapy:** Due to considerable β error, observational studies initially failed to detect significant risks associated with the replacement of endogenous sex steroid hormones in postmenopausal women, leading to widespread use and subsequent exposure to serious health issues (Manson et al., 2013).

Table 1 illustrates how to approach the α–β balance using specific cost figures. We focus on public costs because the example concerns a societal-level decision.

Table 1

Worked Example of Deciding Whether to Implement a Digital Health Application (DHA) to Reduce Pain-Related Disability in Patients With Osteoarthritis

Background	A randomized controlled trial demonstrated a significant effect of a digital health application (DHA) on pain impairment in patients with fibromyalgia, chronic pain and closely related disorders (estimated standardized mean difference, SMD = 0.69, 95% confidence interval = 0.46; 0.92; Etzelmüller et al., 2026). A new study needs to be designed to test the effect for patients with osteoarthritis (OA). How should α and β be chosen according to the costs of these errors?
Costs of the α Error	
DHA is Ineffective but Implemented	According to statutory health insurance, the reimbursed price of introducing DHA amounts to **940** EUR/patient/year. If ineffective, these public resources are wasted. Currently, there is no effective treatment that could be implemented instead of DHA. However, as soon as there is one, implementing the ineffective DHA instead would delay the implementation of this other effective treatment. This would add public health costs that could be prevented by the other treatment. For OA patients with untreated chronic pain, healthcare costs amount to an estimated 6,020 EUR/patient/year, while for patients without pain, they amount to an estimated 3,907 EUR/patient/year (Schild et al., 2021). This results in estimated pain-related excess costs of 2,113 EUR/patient/year. Now, if the other treatment reduces these costs by **c** on a net basis—that is, after accounting for its own expenses, additional costs of **2,113 – c** EUR/patient/year would arise. The total costs amount to **940 + (2,113 - c)** EUR/patient/year.
Costs of the β Error	
DHA is Effective but not Implemented	As above, the pain-related excess costs are assumed to equal 2,113 EUR/patient/year. OA patients with chronic pain deviate by ~1.2 standard deviations (SD) from the healthy population mean in Germany in terms of the quality-of-life instrument SF-36 (Bullinger & Kirchberger, 1998). The

PsychOpen GOLD

	above estimate SMD = 0.69 covers approximately 0.69/1.2 = 58% of this deviation. Hence, not implementing DHA although it is effective, is estimated at 0.58*2,113 = ~**1,215** EUR/patient/year.
Balance	As long as there is no alternative effective treatment, the ratio of the β/α costs equals 1,215 / 940 = **1.29**. For example, with a large desired power to detect the potential treatment effect of 95% (β = 0.05), α should be set to **0.065**. As soon as another effective treatment emerges that reduces the net chronic pain costs by c, the ratio of β/α costs equals 1,215 / (940 + (2,113 − c)). Thus, for c < 1,838, the ratio tips towards α-error costs exceeding β-error costs. For example, if the reduction equals DHA's expenses (c = 940), the ratio becomes 1,215 / 2,113 = **0.58.** With a desired power of 95% (β = 0.05), α should be set to **0.029.** Note that individual costs might be added to these public health cost figures. These relate to reduced quality of life due to chronic pain in terms of 'quality-adjusted life years' (QALYs, He et al., 2026).

In this example, a different assumption may tip the α–β balance in favour of one error type over the other. This underscores the importance of making assumptions explicit and highlights the need to justify them. All examples in this section show that the balance requires justification in each individual case. Moreover, the parsimony argument for favouring a small α can be challenged when an intervention with small individual effects may nevertheless yield considerable cost savings at the societal level (Funder & Ozer, 2019; Greenwald et al., 2015). This argument depends on the scalability of such interventions to the population level, such that their cumulative effects become relevant (Ttofi & Farrington, 2011). However, it seems sufficient to require justification of the *potential* of such interventions, leaving the means of scale-up to politics. Moreover, interventions in fields such as industrial and organizational psychology must account for additional costs, including those associated with productivity, efficiency, and employee engagement.

5. Are the Psychometric Properties of an Instrument Sufficient to Use It? (Confirmatory Research)

Many authors have noted, first, that the behavioural question whether to apply an instrument or not reflects the common purpose underlying psychometric analyses of instruments designed to measure constructs. Second, they argue that the question is often unjustifiably answered in the affirmative (Anvari et al., 2025; Borsboom et al., 2004; Cho, 2021; Flake & Fried, 2020; Hussey & Hughes, 2020; Kline, 2015; McNeish,

PsychOpen GOLD

2018; Scheel et al., 2021; Schimmack, 2021). Specific practices contributing to this overly permissive acceptance include:

- Constructs are measured even though their very existence has not been tested (Borsboom et al., 2004; Michell, 1997).
- Instead of conducting statistical tests, decisions are based solely on point estimates of model fit, validity, and reliability indices.
- Such indices are judged against thresholds for 'acceptable' values. However, what qualifies as 'acceptable' often remains unclear, and thresholds are commonly set so low that they introduce substantial bias into the analysis (Cronbach & Meehl, 1955; Flake et al., 2017; Lash, 2021; Schimmack, 2021).
- Or, for example Cohen's Kappa measuring test-retest reliability is tested against zero, rather than testing whether it exceeds a threshold Δ. A defensible choice of Δ may be the minimum level of reliability at which measurement error introduces only negligible bias into the target analyses.
- Often, only a small sample from a single, very specific population is used, but results are applied to broader, hardly overlapping target populations.

A smaller α raises the bar for introducing a new instrument. We argue that the costs of falsely introducing it (α error) commonly outweighs the costs of falsely dismissing it (β error). In the first case, costs include bias due to poor measurement and ongoing research with instruments and even constructs that lack a solid basis. In the second case, such research is not conducted or is delayed although an appropriate instrument is available. The β error then rather involves the instruments in the lower range of psychometric quality. The practical challenge is that one often has to integrate different kinds of evidence for and against an instrument's usability, and the decision on how to do this is difficult to objectify. However, some general ideas for increasing rigour may still be helpful. A new instrument can be required to pass *each* test that is deemed capable of identifying features of an instrument that would have large costs if used. For example, with two such tests at the common 0.05 level, rigour implies that the overall ('family-wise') α might be as low as 0.025 (if the tests are independent) or as high as 0.05 (if they are perfectly correlated). In our observation, the decision rule, e.g., how a bunch of test results are linked to the overall decision to use an instrument is often vague or adjusted post hoc to permit acceptance.

Given the plethora of existing constructs and instruments (Anvari et al., 2025), accepting one without adequately justified measurement properties is rarely warranted. Exceptions, however, may include urgent or novel contexts, such as emerging public health crises or rapidly evolving technological environments, where the swift development and deployment of new instruments on highly plausible constructs seems warranted. For example, during the COVID-19 pandemic, researchers developed and utilized new scales to assess pandemic-related stress and anxiety (Taylor et al., 2020). Such justifi-

PsychOpen GOLD

cation must include a strong case why using one of the many established instruments (e.g. measuring current stress and anxiety validly and reliably) instead would come at higher costs.

6. Are the Assumptions of a Statistical Method Met so That it May Be Used?

We address this question here because it is used in practice to inform the answers to the previous questions. A wrong answer about the appropriateness of a statistical method increases the probabilities of errors in the decisions on these questions and, if an error actually occurs, the associated costs. In particular, the β for testing whether a statistical method—often a standard test such as the independent samples *t*-test or ANOVA—can be employed is *not chosen for the purpose of assessing a method's applicability*, because it is determined to serve questions 1–5. Consequently, β may only be incidentally appropriate. In small samples, statistical tests for model violations can fail to detect issues like extreme values or outliers, so that these go unnoticed and exert a disproportionate influence on results (Erceg-Hurn & Mirosevich, 2008). Conversely, in very large samples, small deviations from the assumptions are detected—which are often irrelevant (Glass, Peckham, & Sanders, 1972). In both cases, additional, more robust analyses might be carried out. This, however, introduces undesirable variation in confirmatory contexts and invites result-fishing, unless preregistration specifies how to handle the variation (Höfler, 2026). We adopt the suggestion to forgo statistical tests of assumptions in favour of robust alternatives from the outset that rely on these assumptions much less or not at all (Field & Wilcox, 2017; Höfler, 2026; Wilcox, 2017). This approach may avoid both errors at low cost. An example is a test based on robust linear regression instead of ordinary least square regression. It excludes and downweighs extreme values and outliers so that the assumptions of ordinary least squares regression are approximately met (Wilcox, 2017). In general, robust methods should be chosen to withstand foreseeable issues—specifically non-normality, unequal variances, extreme values and outliers. From a Popperian perspective, a hypothesis should hold up under robust testing—that is, it should not pass a test merely because the test relies on faulty assumptions (Höfler, 2026; Popper, 1959).

Table 2 summarizes the arguments for the six questions:

PsychOpen GOLD

Table 2

Informed Decisions and Basic Considerations on the Costs of α and β

What Decision Should Be Informed?	Costs of Type I Error (α)	Costs of Type II Error (β)	Recommendation for Balancing α and β
Should a Hypothesis be Considered Provisionally Accepted? (Confirmatory Research)	Wasted resources, misleading literature, attention spread across too many effects	Missed opportunities, delayed research	Prioritize α if not justified otherwise
Should a Model or Theory Build on the Effect? (Confirmatory Research)	Overfitting, complexity, distraction by spurious effects	Omitted real effects, hindered theoretical progress	Prioritize α unless strong case for small effects
Should a New Hypothesis be Proposed? (Exploratory Research)	Spurious patterns, communication overload, reduced credibility	Missed early insights, delayed progress	Prioritize α, the more, the more exploration has been used, allow lower β if patterns may be supported by theory.
Should a New Intervention be Implemented? (Confirmatory Research)	Ineffective treatment, unjustified adverse effect, wasted resources	Missed benefits, delayed intervention with individual and public health consequences	Case-by-case justification; α or β may dominate depending on an intervention's prior plausibility and adverse effects
Are the Psychometric Properties of an Instrument Sufficient to Use it? (Confirmatory Research)	Flawed constructs, erroneous measures, biased research	Delayed use of acceptable instruments	Generally prefer lower α, tolerate higher β
Are the Assumptions of a Statistical Method Met so That it may be Used?	Inappropriate method, spurious results due to wrong assumptions	Unnecessary variation in results by avoiding usable tests	Minimize both errors by using robust methods instead of standard methods and assumption testing

Overarching Aspects

Several other issues apply across all six decisional questions or interconnect them. Key insights on these can be discussed at a general, overarching level.

PsychOpen GOLD

Different Decisions With the Same Test

In addition to questions 1-5, the same test can even *inform different behavioural decisions* based on the same data, and *different α levels* might be justified by the distinct consequences of each decision. Thus, the same p-value might lead to divergent actions, for example, provisionally accepting an effect but not (yet) conducting a costly intervention study based on it. Aisbett (2024) gives the example of Molnupiravir treatment for COVID-19, where different decisions may arise depending on which types of costs one aims to minimize, purely economic burden or also individual suffering of patients. When planning a study, the sample size must be large enough to achieve the required β for *each* individual decision, therefore, the *smallest* β is decisive.

Multiple Testing and Overall Decisions

In decisions 1–6, hypotheses are often evaluated using multiple statistical tests. This requires an overall decision to be made based on the set of p-values obtained from the tests. Greenland (2021) recommends combining tests in a way that ensures the chosen α and β levels are adhered to. This approach indicates whether the α level for individual tests should be adjusted for the number of tests, k, most simply via the Bonferroni correction (dividing the α for each individual test by k) when an OR-decision—rejecting the global hypothesis if at least one individual test is significant—is applied. Psychological research often claims a common context that justifies such adjustment, but in our experience the arguments for such a context are rarely explicated. For instance, testing whether 'Psychotherapy A is superior to a waiting list' and 'Psychotherapy B is superior to a treatment as usual' involves separate hypotheses with two distinct decisions: Should A be implemented, and should B be implemented? Here, correction for multiple testing would penalize each test just due to the presence of the other, leading to an undesirably low α and, consequently, an undesirably high β (Bender & Lange, 2001; Greenland, 2021; Rubin, 2024).

Is a New Study Required Because the Current Study is Inconclusive on the Question of Interest?

The question concerns datasets without thorough sample size planning and secondary data analyses where the sample size is insufficient to achieve the desired β. Here, the answer is commonly 'yes.' Instead, researchers take 'unclear' or 'mixed evidence' to behave as if an effect was truly corroborated, for example, because they lack time and money to conduct another study. With this practice, however, the desired α is exceeded.

The need for new studies is a much-debated topic in interpreting post hoc analyses of certain treatments for depression (Munder et al., 2019). A notable example is electroconvulsive therapy, which, though established, remains controversial. When testing its efficacy with a given data set to decide whether new randomized controlled trials are

PsychOpen GOLD

needed, a low α entails ethical costs—subjecting patients to a potentially ineffective sham intervention. Conversely, the negative results of underpowered studies constitute a poor justification for the continued use of a treatment that carries potential harm (Read, 2022).

Discussion

We begin by noting fundamental limits of what statistics can accomplish and instances in which statistical tests may unnecessarily introduce errors into decisions. Statistics is not an end in itself but a toolbox to help answer substantive questions in empirical science (Gigerenzer, 2004; Greenland, 2019; Lakens et al., 2018a; Meehl, 1978). Statistical inference offers a wide range of methods and specifications—a complexity that can easily overwhelm non-experts and opens the door to fish for desired results. However, translating statistical options (such as the choice of α and β) into substantive considerations—like the costs of making incorrect decisions—facilitates unavoidable choices or, at least, prompts a much-needed discussion of the underlying assumptions (MacCallum et al., 1992). As argued, in confirmatory research, substantive knowledge should guide model specification and override data-driven selection, which can produce spurious associations, overfitting, inflated α, and reduced reproducibility (Babyak, 2004; MacCallum et al., 1992; Thompson, 1995; Wysocki et al., 2022). Theory-based model building comes at the acceptable or even desirable cost of conceptual rigour and avoids the costs of adding further error into decisions. Other examples of the misguided use of statistical tests include:

- *Testing Coefficients of Confounders*, which should be included in a model because they are identified a priori as common causes of both exposure and outcome. Such variables should be included based on theoretical justification, not on statistical significance, to avoid omitted variable bias (Greenland, 2008; Wysocki et al., 2022).
- *Testing Baseline Covariates in Randomized Experiments:* Statistical tests are intended to distinguish systematic effects from chance variation. In randomized studies, group balance is achieved on average through randomization. Any observed imbalances are due to chance, not bias, and thus do not require statistical testing (Senn, 1994).

Thus, the widely discussed theory crisis in psychology (Meehl, 1978; Gigerenzer, 2004; Oberauer & Lewandowsky, 2019) is accompanied by an overreliance on statistics. If psychology generally confines its theories and predictions to large effects, preferring small α over small β, it could more easily accommodate simpler statistical models that require fewer assumptions. Conversely, whenever researchers struggle to make statistical decisions on substantive grounds, parsimony can serve as a rational fallback. The provided example illustrates how explicit assumptions and their relation to the result can replace arbitrariness and encourage refinement of those assumptions. While some subjectivity in assessing error costs is generally inevitable, better-informed decisions are

PsychOpen GOLD

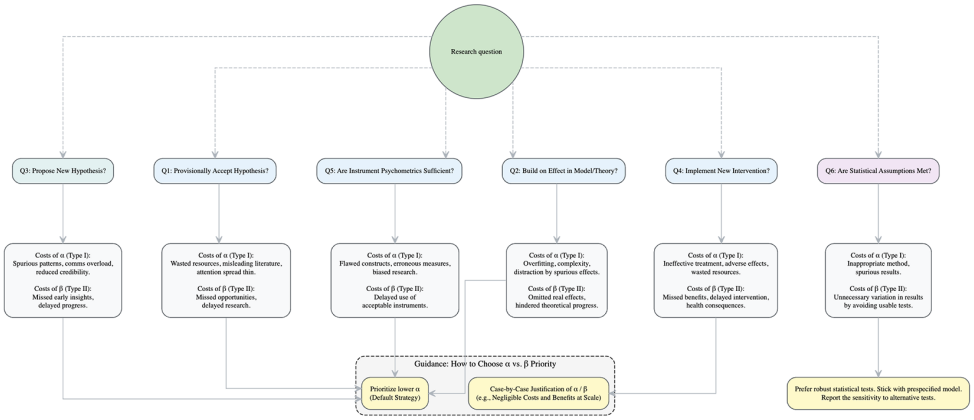
often possible and clearly preferable to the common practice of making such decisions without any justification. The better informed a justification, the more likely the decision is to gain acceptance; the weaker it is, the more it invites critical scrutiny and the need for refinement.

Current justifications for preferences of low β over small α include situations where interventions pose low risks (e.g. new treatments with minimal adverse effects) or where small effects may truly accumulate to yield considerable societal benefits. We have seen that some of these exceptions require more in-depth argumentation, such as whether an aggregate of large effects constitutes a meaningful unit and construct for explanation and prediction of behaviour. Argumentation must also consider the different levels of error cost beyond the type of decision-making question: the research field and the specific context. Open discussion of such issues always provokes further arguments and counterarguments with explicit assumptions just as what science is and research should be.

We would like to conclude with a flowchart that summarises our considerations. Figure 2 can be used by researchers to guide their assessment of error costs. The chart can also serve as a basis for future elaboration in specific fields of psychology and contexts, inviting refinement, more detail and concrete cost calculations.

Figure 2

Flowchart Summarizing the Error Balance of Statistical Tests Used to Address Different Research Questions



PsychOpen GOLD

Funding: The authors have no funding to report.

Acknowledgments: We wish to thank Wendy Higgins for valuable hints on psychometric testing.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: Michael Höfler conceived the basic idea for the article and took the lead in writing. Sebastian Trautmann contributed to the structure, writing, and development of substantive examples. Robert Miller provided input on statistical issues and added clinical examples, including the cost calculation for the burden of chronic pain in osteoarthritis patients in Germany. Philipp Kanske contributed to the writing.

References

Aisbett, J. (2024). Can expected error costs justify testing a hypothesis at multiple alpha levels rather than searching for an elusive optimal alpha? *PLoS One, 19*(9), Article e0304675. https://doi.org/10.1371/journal.pone.0304675

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *BMJ (Clinical Research Ed.), 311*(7003), Article 485. https://doi.org/10.1136/bmj.311.7003.485

Anvari, F., Alsalti, T., Oehler, L. A., Hussey, I., Elson, M., & Arslan, R. C. (2025). Defragmenting psychology. *Nature Human Behaviour, 9*, 836–839. https://doi.org/10.1038/s41562-025-02138-0

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411–421. https://doi.org/10.1097/01.psy.0000127692.23278.a9

Bender, R., & Lange, S. (2001). Adjusting for multiple testing--when and how? *Journal of Clinical Epidemiology, 54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). Springer.

Bonander, C., Ekman, M., & Jakobsson, N. (2022). Vaccination nudges: A study of pre-booked COVID-19 vaccinations in Sweden. *Social Science & Medicine, 309*, Article 115248. https://doi.org/10.1016/j.socscimed.2022.115248

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

PsychOpen GOLD

Bullinger, M., & Kirchberger, I. (1998). *SF-36 Fragebogen zum Gesundheitszustand: Handanweisung*. Hogrefe.

Cartwright, N., Hardie, J., Montuschi, E., Soleiman, M., & Thresher, A. C. (2022). *The tangle of science: Reliability beyond method, rigour, and objectivity*. Oxford University Press.

Cho, E. (2021). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods, 24*(2), 630–667.

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

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Dawid, P. (2021). Decision-theoretic foundations for statistical causality. *Journal of Causal Inference, 9*(1), 39–77. https://doi.org/10.1515/jci-2020-0008

DeGroot, M. H. (1970). *Optimal statistical decisions*. McGraw-Hill.

Duhem, P. (1954). *The aim and structure of physical theory* (P. P. Wiener, Trans.). Princeton University Press. (Original work published 1906)

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *The American Psychologist, 63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Etzelmüller, A., Barke, A., & Feiler, M. (2026). Digital acceptance and commitment therapy for the treatment of chronic pain: Findings of a randomized controlled efficacy trial. *Deutsches Ärzteblatt International, 123*, 149–156. https://doi.org/10.3238/arztebl.m2025.0239

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.

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. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE, 9*(10), Article e109019. https://doi.org/10.1371/journal.pone.0109019

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Furukawa, T. A., Cipriani, A., Atkinson, L. Z., Leucht, S., Ogawa, Y., Takeshima, N., Hayasaka, Y., Chaimani, A., & Salanti, G. (2016). Placebo response rates in antidepressant trials: A systematic

PsychOpen GOLD

review of published and unpublished studies. *The Lancet Psychiatry, 3*(11), 1059–1066. https://doi.org/10.1016/S2215-0366(16)30307-8

Gelfand, A. E., & Wang, F. (2002). A simulation-based approach to sample size determination under a given model and for separating models. *Statistical Science, 17*(2), 193–208. https://doi.org/10.1214/ss/1030550861

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

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Sage Publications.

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

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237–288. https://doi.org/10.3102/00346543042003237

Goeman, J. J., Solari, A., & Stijnen, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine, 29*(20), 2117–2125. https://doi.org/10.1002/sim.4002

Gordon, J. A. (2021). *F*rom small effects to big impacts [Editorial]. *Complex Psychiatry, 7*(1–2), 9–10. https://doi.org/10.1159/000517267

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science, 17*(1), 205–215. https://doi.org/10.1177/1745691620984483

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 168*(2), 267–306. https://doi.org/10.1111/j.1467-985X.2004.00349.x

Greenland, S. (2008). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology, 167*(5), 523–531. https://doi.org/10.1093/aje/kwm355

Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician, 73*(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Greenland, S. (2021). Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. *Paediatric and Perinatal Epidemiology, 35*(1), 8–23. https://doi.org/10.1111/ppe.12711

Greenland, S. (2022). The need for cognitive science in methodology. *American Journal of Epidemiology, 191*(8), 1351–1356. https://doi.org/10.1093/aje/kwac037

PsychOpen GOLD

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology, 108*(4), 553–561. https://doi.org/10.1037/pspa0000016

He, J., Wei, H., Mo, Q., Liu, Y., Li, X., Yang, Y., & Tian, L. (2026). A comparative analysis of health technology assessment decision-making for cancer drugs in Germany and England. *Frontiers in Public Health, 14*, Article 1733239. https://doi.org/10.3389/fpubh.2026.1733239

Höfler, M. (2026). Robust tests should be the default, not the backup. *Peer Community Journal, 6*, Article e670. https://doi.org/10.24072/pcjournal.670

Höfler, M., Scherbaum, S., Kanske, P., McDonald, B., & Miller, R. (2022). Means to valuable exploration I: The blending of confirmation and exploration and how to resolve it. *Meta-Psychology, 6*. https://doi.org/10.15626/MP.2021.2837

Höfler, M., Kanske, P., McDonald, B., & Miller, R. (2023). Means to valuable exploration: II. How to explore data to modify existing claims and create new ones. *Meta-Psychology, 7*. https://doi.org/10.15626/MP.2022.3270

Hussey, I., & Hughes, S. (2020). Hidden invalidity: Fifteen years of insufficient validity reporting in the *Journal of Positive Psychology. Advances in Methods and Practices in Psychological Science, 3*(2), 166–184. https://doi.org/10.1177/2515245919882903

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

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review, 13*(4), 549–562. https://doi.org/10.3758/BF03193962

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759. https://doi.org/10.1177/0013164496056005002

Kline, P. (2015). *Handbook of psychological testing* (2nd ed.). Routledge. https://doi.org/10.4324/9781315812274

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S. C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., . . .Zwaan, R. A. (2018a). Justify your alpha. *Nature Human Behaviour, 2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

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

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports. *Evidence-Based Toxicology, 2*(1), . Article 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Lakens, D., Scheel, A. M., & Isager, P. M. (2018b). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lash, T. L. (2021). Bias analysis. In T. L. Lash, T. J. VanderWeele, S. Haneuse, & K. J. Rothman (Eds.), *Modern epidemiology* (4th ed., pp. 345–380). Wolters Kluwer.

Lindley, D. V. (1985). *Making decisions* (2nd ed.). Wiley.

PsychOpen GOLD

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

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

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Manson, J. E., Chlebowski, R. T., Stefanick, M. L., Aragaki, A. K., Rossouw, J. E., Prentice, R. L., Anderson, G., Howard, B. V., Thomson, C. A., LaCroix, A. Z., Wactawski-Wende, J., Jackson, R. D., Limacher, M., Margolis, K. L., Wassertheil-Smoller, S., Beresford, S. A., Cauley, J. A., Eaton, C. B., Gass, M., …Wallace, R. B. (2013). Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. *Journal of the American Medical Association, 310*(13), 1353–1368. https://doi.org/10.1001/jama.2013.278040

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge University Press

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Munder, T., Flückiger, C., Leichsenring, F., Abbass, A. A., Hilsenroth, M. J., Luyten, P., Rabung, S., Steinert, C., & Wampold, B. E. (2019). Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiology and Psychiatric Sciences, 28*(3), 268–274. https://doi.org/10.1017/S2045796018000355

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *The Journal of Applied Psychology, 84*(2), 234–248. https://doi.org/10.1037/0021-9010.84.2.234

Nagy, T., Hergert, J., Elsherif, M. M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., Wang, Y. A., Scharfenberg, D., Tyson, G., Yang, Y. F., Skvortsova, A., Alarie, S.,

PsychOpen GOLD

Graves, K., Sotola, L. K., Moreau, D., & Rubínová, E. (2025). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science, 8*(3), 1–36. https://doi.org/10.1177/25152459251348431

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Pearson, E. S. (1962). Some thoughts on statistical inference. *Annals of Mathematical Statistics, 33*, 394–403. https://doi.org/10.1214/aoms/1177704566

Petersen, I. T. (2024). *Principles of psychological assessment: With applied examples in R* (1st ed). Taylor & Francis. https://doi.org/10.1201/9781003357421-9

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson & Co.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review, 60*(1), 20–43. https://doi.org/10.2307/2181906

Raiffa, H., & Schlaifer, R. (1961). *Applied statistical decision theory*. Division of Research, Graduate School of Business Administration, Harvard University.

Rafi, Z., & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: Replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology, 20*, Article 244. https://doi.org/10.1186/s12874-020-01105-9

Read, J. (2022). A response to yet another defence of ECT in the absence of robust efficacy and safety evidence. *Epidemiology and Psychiatric Sciences, 31*, Article e13. https://doi.org/10.1017/S2045796021000846

Robert, C. P. (2007). *The Bayesian choice: From decision-theoretic foundations to computational implementation* (2nd ed.). Springer.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Ross, E. L., Vijan, S., Miller, E. M., Valenstein, M., & Zivin, K. (2019). The cost-effectiveness of Cognitive Behavioural Therapy versus second-generation antidepressants for initial treatment of major depressive disorder in the United States: A decision analytic model. *Annals of Internal Medicine, 171*(11), 785–795. https://doi.org/10.7326/M18-1480

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual. *Methods in Psychology, 10*, Article 100140. https://doi.org/10.1016/j.metip.2024.100140

Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2). https://doi.org/10.1177/25152459211007467

Schild, M., Müller, U., von Schenck, U., Prieur, S., & Miller, R. (2021). The burden of chronic pain for patients with osteoarthritis in Germany: A retrospective cohort study of claims data. *BMC Musculoskeletal Disorders, 22*(1), Article 317. https://doi.org/10.1186/s12891-021-04180-1

PsychOpen GOLD

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology, 5*, Article 1645. https://doi.org/10.15626/MP.2019.1645

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics, 102*(1), 411–432. https://doi.org/10.1007/s11192-014-1251-5

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *Lancet, 365*(9467), 1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3

Schwaba, T., Clapp Sullivan, M. L., & Akingbuwa, W. A. (2025). *Robust inference and widespread genetic correlates from a large-scale genetic association study of human personality*. bioRxiv. https://doi.org/10.1101/2025.05.16.648988v1

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine, 13*(17), 1715–1726. https://doi.org/10.1002/sim.4780131703

Spanos, A. (2013). Who should be afraid of the Jeffreys-Lindley Paradox? *Philosophy of Science, 80*(1), 73–93. https://doi.org/10.1086/668875

Sugarman, M. A. (2016). Are antidepressants and psychotherapy equally effective in treating depression? A critical commentary. *Journal of Mental Health, 25*(6), 475–478. https://doi.org/10.3109/09638237.2016.1139071

Taylor, S., Landry, C. A., Paluszek, M. M., Fergus, T. A., McKay, D., & Asmundson, G. J. G. (2020). Development and initial validation of the COVID Stress Scales. *Journal of Anxiety Disorders, 72*, Article 102232. https://doi.org/10.1016/j.janxdis.2020.102232

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guideline editorial. *Educational and Psychological Measurement, 55*(4), 525–534. https://doi.org/10.1177/0013164495055004001

Ttofi, M. M., & Farrington, D. P. (2011). Effectiveness of school-based programs to reduce bullying: A systematic and meta-analytic review. *Journal of Experimental Criminology, 7*(1), 27–56. https://doi.org/10.1007/s11292-010-9109-1

Uygun Tunç, D., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology, 33*(3), 403–423. https://doi.org/10.1177/09593543231160112

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2018). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.

Wald, A. (1950). *Statistical decision functions*. Wiley.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Westra, H. A., Arkowitz, H., & Dozois, D. J. (2009). Adding a motivational interviewing pretreatment to cognitive behavioral therapy for generalized anxiety disorder: A preliminary randomized controlled trial. *Journal of Anxiety Disorders, 23*(8), 1106–1117. https://doi.org/10.1016/j.janxdis.2009.07.014

PsychOpen GOLD

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Academic Press.

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical control requires causal justification. *Advances in Methods and Practices in Psychological Science, 5*(2). https://doi.org/10.1177/25152459221095823

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

About the Authors

Michael Höfler is a methodologist in quantitative psychology. His work focuses on metascience, statistical methods, rigor, transparency in assumptions, and causal inference.

Sebastian Trautmann is a clinical psychologist. He investigates the effects of stress and traumatic events on the development and the treatment of mental disorders.

Philipp Kanske is a clinical psychologist and neuroscientist whose research focuses on social cognition, emotion regulation, and mental health. He investigates the neural and psychological mechanisms underlying empathy and social behaviour.

Robert Miller is a methodologist and researcher in psychology. His work addresses stress, digital health applications, modeling of feature distributions and changes, clinical outcome assessment, and signal detection in high-dimensional data.

PsychOpen GOLD is a publishing service by
Leibniz Institute for Psychology (ZPID), Germany.
www.leibniz-psychology.org