aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2024, Vol. 19(1) 223–243

DOI: 10.1177/17456916231182097
www.psychologicalscience.org/PPS

Sage

Better Accuracy for Better Science . . . Through Random Conclusions

Clintin P. Davis-Stober[1], Jason Dana[2], David Kellen[3], Sara D. McMullin[4], and Wes Bonifay[5]

[1]Department of Psychological Sciences, MU Institute for Data Science and Informatics, University of Missouri; [2]Yale School of Management, Yale University; [3]Department of Psychology, Syracuse University; [4]Department of Psychological Sciences, University of Missouri; and [5]Missouri Prevention Science Institute, Educational, School & Counseling Psychology, University of Missouri

Abstract

Conducting research with human subjects can be difficult because of limited sample sizes and small empirical effects. We demonstrate that this problem can yield patterns of results that are practically indistinguishable from flipping a coin to determine the direction of treatment effects. We use this idea of random conclusions to establish a baseline for interpreting effect-size estimates, in turn producing more stringent thresholds for hypothesis testing and for statistical-power calculations. An examination of recent meta-analyses in psychology, neuroscience, and medicine confirms that, even if all considered effects are real, results involving small effects are indeed indistinguishable from random conclusions.

Keywords

random conclusions, estimation, hypothesis testing, *t* tests, benchmarks

Introduction

Human-subjects research often involves noisy measures and limited sample sizes. Accordingly, small effects and low statistical power are typical in many areas of behavioral and medical science (Marek et al., 2022; Szucs & Ioannidis, 2017). Some argue that this situation is tenable because the ongoing identification of small effects amounts to a steady accumulation of knowledge (Götz et al., 2022). We argue to the contrary. Specifically, we show that the study of small effects frequently produces results that are indistinguishable from flipping a coin to determine the direction of an experimental treatment's effect. We use this idea to develop a benchmark based on minimum acceptable estimation accuracy. This benchmark yields an intuitive interpretation of effect-size estimates—one based in accurate estimation. We show that calibrating existing tests to our benchmark yields far stricter thresholds for hypothesis testing and for statistical-power calculations. Our work is intended to spark a larger discussion within the scientific community on acceptable estimation accuracy, the interpretation of effects, and statistical standards.

Although there are many exceptions, behavioral scientists almost universally test null hypotheses, which are often formulated as two or more means being exactly equal to one another. Much ink has been spilled noting the shortcomings of this approach (e.g., Krantz, 1999; Nickerson, 2000; van de Schoot et al., 2011). Cohen (1994) famously criticized the null hypothesis through his "nil" hypothesis critique, describing it as a conceptual tool that is ill-suited for answering substantive research questions. He noted that for continuous dependent variables, it is simply impossible for two population means to be truly equal to one another. This means that the null hypothesis acts as a straw man to be knocked down at a given sample size. By his critique, all effects exist in a trivial sense; it just may be that some are so small that they do not warrant attention. A more meaningful line of investigation is

Corresponding Author:
Clintin P. Davis-Stober, University of Missouri, Department of Psychological Sciences
Email: stoberc@missouri.edu

determining whether effects are accurately estimated and characterized.

What constitutes acceptable estimation accuracy? This question is challenging to answer and fraught with subjectivity. A confidence interval (CI) deemed acceptably narrow by one scientist may be unacceptably wide to another. We seek to answer this question by the use of a reference—a foil—with undeniable negative qualities. To better understand the accuracy of standard methods, we will compare them against a foil estimation process that is, by construction, incapable of accurately estimating effects. Such a foil is useful for handling questions of subjectivity. If a community of scientists agree that this foil is unacceptably inaccurate, then any estimation process that cannot be distinguished from it is also unacceptably inaccurate.

Our foil must be tailored to the types of questions that behavioral scientists ask and to how they make decisions about data. Behavioral scientists often formulate directional hypotheses about treatment effects. Is the population mean of Group A larger than that of Group B? A strong foil would offer zero information about the correct direction of effects. A foil could randomize the direction of any observed effect; for example, which group mean is larger than another would be decided via a coin flip. Such a foil creates a worst-case scenario for evaluating any directional hypothesis. In addition, behavioral scientists typically use the outcome of a statistical test to conclude whether a treatment effect is detected. In keeping with our estimation focus, an ideal foil would remove effect detection from the comparison. One way to handle this is for the foil to correctly detect whether an effect exists at similar, or identical, rates as standard methods. A scientist using this foil would correctly reject a relevant null hypothesis just as often as someone using standard estimation methods. This would make the foil especially useful for evaluating published findings in the literature.

Scientists using such a foil would arrive at random conclusions regarding their data. All else being equal, they would detect effects as often as scientists using standard methods, but would be incapable of accurately estimating and characterizing them. The logic is straightforward: If one accepts that arriving at random conclusions is unscientific and inaccurate, then it becomes incumbent on the scientific community to use statistical procedures that would be distinguishable from such a foil.[1] In the present work, we focus on the canonical case of using sample means to estimate population means for two independent groups. Our proposed foil consists of an estimation process that randomizes the direction of treatment effects while still correctly rejecting a null hypothesis as often as standard methods.

Our analyses reveal that distinguishing sample means from such a foil requires far larger sample sizes than typically employed in the behavioral sciences, especially when studying the kinds of small effects that are commonplace in the psychological literature. We also show that our foil comparison naturally relates to many existing tests and methods, including those based on traditional null hypotheses. We leverage these connections to provide new calibrations for existing techniques. For power analyses, we show that typical power thresholds of .80 are not sufficient to rule out unacceptable estimation accuracy. Linking our argument to hypothesis testing, we show that far stricter thresholds ($\alpha = .0005$) are required if sufficient estimation accuracy is to be ensured. We also provide a simple methodology that allows researchers to convert a common measure of effect size, Cohen's d, into an easily understood measure of estimation accuracy on the basis of our foil. This methodology can be applied to CIs over Cohen's d, allowing researchers to determine whether their estimates are acceptably accurate. Finally, we examine a collection of meta-analyses from the behavioral sciences, finding that typical estimates in many fields of study are indistinguishable from our random conclusions foil.

Ultimately, all scientific decisions regarding data are made by human beings. A key aim of any statistical methodology is to provide characterizations of data that researchers can understand. What we provide in the current work is simply a perspective, one grounded in a common experimental design with linkages to many other familiar statistical quantities and methods. It is through this framing that we aim to push forward the conversation on estimation accuracy and replication efforts. To further understand our approach and provide precise definitions, consider the following scenario.

A Tale of Two Labs

Consider two hypothetical laboratories, Lab 1 and Lab 2, studying an effect—for instance, the efficacy of a drug. Both labs use a treatment condition (Group A) and a control condition (Group B) and compare the sample means from each group, $\overline{x}_A$ and $\overline{x}_B$, on some outcome measure. These sample means underpin the statistical tests conducted by both labs and provide point estimates for the population means, μ_A and μ_B, that instantiate their scientific hypotheses regarding the drug's effect. Assume that the drug has a true effect $\delta > 0$, where $\delta = \frac{\mu_A - \mu_B}{\sigma}$, with σ being the standard deviation of responses from the populations.[2]

Unfortunately, Lab 2 has a glitch in their data-analysis software—it randomly assigns, with equal likelihood,

the labels of "treatment" and "control" to those means. That is, if Lab 2 conducted a study for which the actual sample means for the two conditions were $\bar{x}_A = 7$ and $\bar{x}_B = 3$, the software would instead report $\bar{x}_A = 3$ and $\bar{x}_B = 7$ with probability equal to .5, and the truth cannot be recovered. We refer to this procedure as a *random-conclusions estimator* (RCE) because the direction of the effect—whether the drug helps or harms—is determined at random. Although mathematically related, the RCE is distinct from a classic Fisher randomization test in which labels are randomized at the individual response level to generate a null, no-effect reference distribution.

If Lab 2's error came to light, retraction of any study that relied on this software would be demanded, and a drug approved on the basis of such results would (rightfully) be recalled. But Lab 2 provides an interesting comparison with Lab 1, especially when considering issues of replication and reliability. Lab 2 will correctly reject the null hypothesis, $H_0 : \mu_A = \mu_B$, exactly as often as Lab 1 using a two-tailed *t* test. Barring preregistration restrictions, both labs will publish results at similar rates. In this way, Lab 2 will pollute the scientific literature with random conclusions and, in the case of drug trials, potentially claim evidence for dangerous treatments.

Lab 1 and Lab 2 are identical with the exception that Lab 2 is using an RCE, which, by any measure, is not science because the direction of effects (including published effects) is determined via a coin flip. Intuitively, we would like to believe that results from the two labs would be readily distinguishable. Unfortunately, in many areas of behavioral science, even if all effects exist, Lab 2's results will often be strikingly similar to Lab 1's, and the gain from removing their results from the literature may be marginal at best. This situation is illustrated in Figure 1, which presents scenarios for effect sizes that are conventionally considered large, medium, and small (yet interpretable; Cohen, 1988; Sawilowsky, 2009). For simplicity, these scenarios assume that outcomes in both conditions are normally distributed with unit variance. The left and right columns of Figure 1 illustrate the sampling distributions of mean estimates in each of the labs. Each dot represents a pair of means from a single study. How well these means estimate the population means μ_A and μ_B is quantified in terms of a common metric for assessing estimation accuracy: mean-squared error (MSE; see the Appendix).

In the top row of Figure 1, the effect size is large. Lab 2's bimodal distribution of estimates clearly evidences the software error, and the resulting MSE is 19 times larger than Lab 1's. We use ψ to denote the ratio $\frac{\text{MSE}_{\text{Lab2}}}{\text{MSE}_{\text{Lab1}}}$; ψ has a lower bound of 1, given that there is no scenario in which Lab 2's estimates will be, on average, more accurate than Lab 1's. The middle and bottom rows of Figure 1 illustrate how the estimates from the two labs converge as effect size becomes smaller, with Lab 2's distribution of estimates eventually becoming unimodal. These changes are indexed by ψ: In the bottom row, $\psi = 1.5$, and estimates from the two labs are visually nearly indistinguishable, an impression confirmed by a small Wasserstein metric (Rubner et al., 2000) and the large number of replicates needed (at least 54 per lab) to reliably distinguish the distributions of results from the two labs via a Kolmogorov-Smirnov test (see the Appendix).

Effect size and sample size combinations like those in the bottom row of Figure 1 raise an important question: If Lab 2's results are subject to retraction, how should we interpret Lab 1's results? Put differently, if one's results look unscientific, perhaps they are unscientific. A computer glitch on the scale of Lab 2's results is, one hopes, an unlikely occurrence, but the comparison is useful in illustrating what a worst-case estimator could look like and why it would be problematic if it were indistinguishable from current practice. Within the behavioral sciences, many of the hypotheses being tested, if not the vast majority, are directional in nature. The RCE completely randomizes the direction of effects, removing any information about direction from the data. Yet the RCE is special in that it still detects effects at the same rate as sample means via a nondirectional *t* test, which is, once again, ubiquitous practice in the behavioral sciences. In this way, our RCE comparison provides an interesting new perspective on published literature in the field, which often hinges upon the successful reporting of a significant test. We are not seriously suggesting that such a computer glitch exists, but we do think it highly problematic if a large corpus of work within the behavioral sciences is indistinguishable from such an error.[3]

General Formulation

If the goal is to be distinguishable from a veritable Lab 2, as instantiated by the RCE, we can use ψ as an index to set standards for hypothesis testing and sample-size planning. As shown in the Appendix, ψ simplifies to

$$\psi(\delta, n) = \frac{n\delta^2 + 2}{2}, \tag{1}$$

where n is the sample size per group. Equation 1 is straightforward to interpret: For given values of δ and n, sample means are ψ times as accurate (in terms of MSE) as the RCE. Although ψ is distribution-free and interpretable outside of any testing framework, it functionally relates to a two-sample *t* test and the resulting *p* values. See the Appendix for connections between ψ

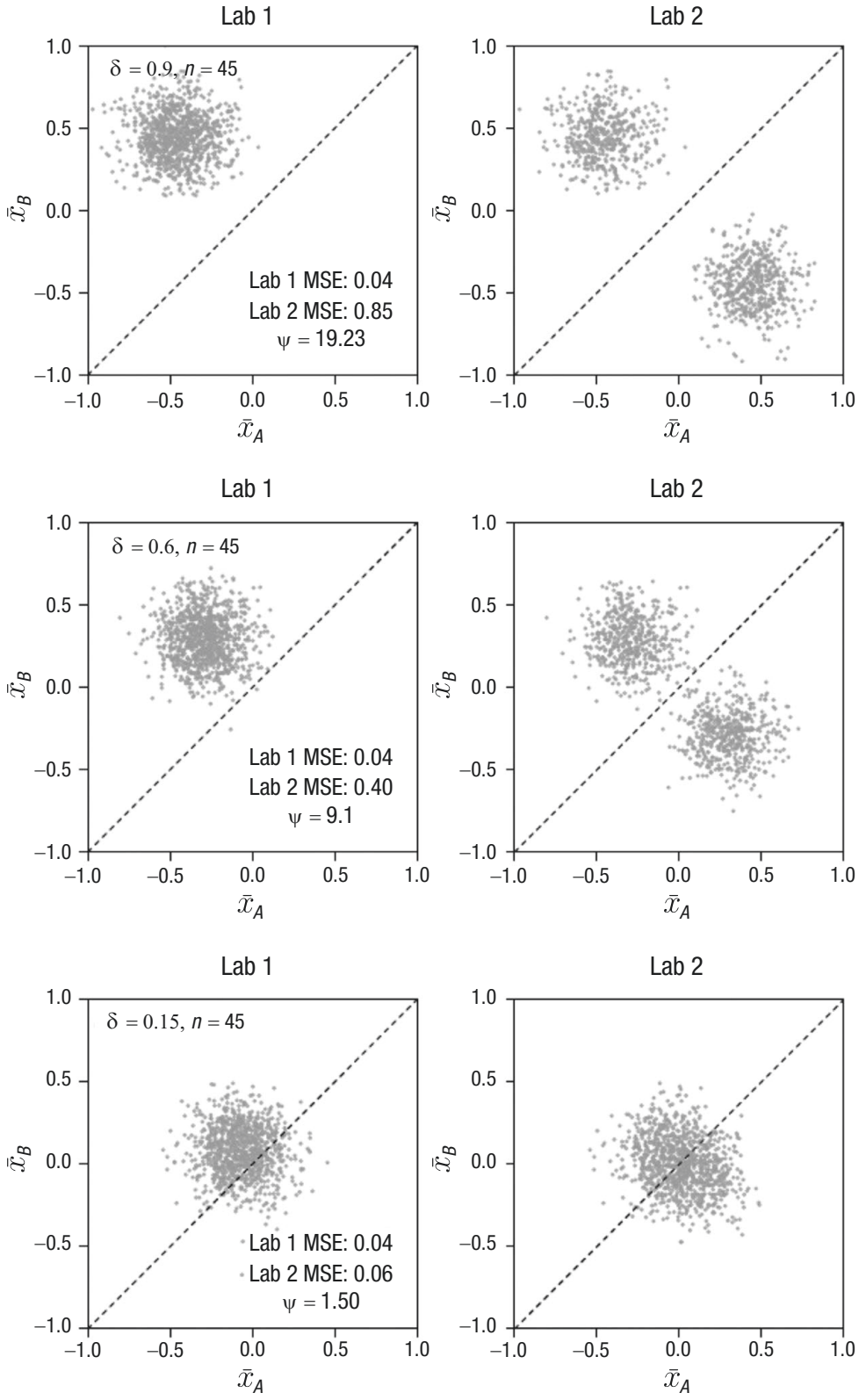

Fig. 1. Distribution of sample mean estimates $\bar{x}_A$ and $\bar{x}_B$ for Lab 1 and Lab 2. Each row corresponds to a different combination of effect size d and sample size per group n. The ratio of mean-squared error (MSE) values for the two labs, $\frac{\text{MSE}_{\text{Lab2}}}{\text{MSE}_{\text{Lab1}}}$, is represented by ψ. To facilitate visualization, we report all relevant values for each comparison from both Lab 1 and Lab 2 (δ, n, MSE) in the Lab 1 panel.

and other metrics, including out-of-sample R^2. This relationship allows us to reexamine hypothesis-testing and statistical-power standards by calibrating to minimally acceptable estimation, as opposed to detection error rates against a null hypothesis. The mathematics are familiar, but the RCE comparison offers new interpretation to these techniques.

Determining a minimum acceptable ψ for a given scientific discipline is perhaps best decided on a case-by-case basis, taking into consideration specific research goals (S. F. Anderson & Maxwell, 2016; Navarro, 2019). Here, we demonstrate the consequences of a threshold of 3 for the interpretation of results and sample-size planning. Although somewhat arbitrary and perhaps modest, this threshold is motivated by the logic illustrated in Figure 1. When $\psi < 3$, the sampling distribution of the RCE becomes unimodal for normal random variables (Figs. A3–A7, Appendix), and the number of study replicates required to reliably distinguish it from sample means becomes impractical (Table A1). If we take our illustration with the two labs seriously, poor ψ values imply that members of Lab 1 and Lab 2 could spend their entire careers replicating scores of studies and be unable to reject the null hypothesis that they are using the same estimator (see the Appendix).

Table A1 in the Appendix characterizes ψ in terms of the information about the direction of effect that is gained by using sample means versus the RCE. For example, for $\psi = 1.5$, the usage of sample means reduces the uncertainty about the correct direction of effects by only 29% compared with the total uncertainty given by the RCE (see also Fig. A1, Appendix). In this way, our RCE comparison links directly to the concept of Type S errors regarding the sign of the effect (Gelman & Carlin, 2014; Gelman & Tuerlinckx, 2000). See also recent work by Domingue et al. (2021), who applied the concept of weighted coins to develop a measure of predictive accuracy for binary outcomes.

Applications to CIs and Hypothesis Testing

Applying Equation 1 to the bounds of a 95% CI over δ provides researchers a simple, transparent method to gauge how accurately a range of plausible effects is being estimated. For example, consider a study with a sample size of 50 that yields an effect size point estimate d of 0.5 and a 95% CI equal to [0.10, 0.89] (see, e.g., Cumming & Finch, 2001). This interval does not include 0, corresponds to a p value of .014, and by current standards would provide researchers assurance that an effect has been detected. But even if this interval contains the population value δ, researchers cannot be confident that their estimation is better than the bottom row of Figure 1. Applying Equation 1 to this CI yields a ψ interval of [1.25, 20.80], which includes conditions in which sample mean estimates are practically indistinguishable from random conclusions. Put another way, these researchers may claim that the population means are not equal, but, upon examining the bounds on ψ, may also conclude that there remains tremendous uncertainty regarding the size and direction of the effect. Indeed, sample means estimation yields a 16.075% reduction in uncertainty (relative to the RCE) at the lower bound ($\psi = 1.25$) and a 99.998% reduction in uncertainty at the upper bound ($\psi = 20.80$; Fig. A1). Although the effect-size estimate implies a difference between groups, the accuracy of this estimate could be anything from a blind guess to a statement of fact.

Figure 2 contextualizes ψ within familiar statistical quantities:

- Panel (a) - ensuring that ψ is greater than 3 often requires a large n, especially when dealing with smaller effect sizes (e.g., $\delta = 0.10$). Sample size requirements are more stringent if one also wants to achieve 95% confidence that the true ψ is larger than 3. For example, the estimation accuracy of a small effect (with $\delta = 0.3$) requires a sample of $2 \times 255 = 510$ to be confidently acceptable. See the Appendix for a discussion on how effect-size priors can be used to determine n.
- Panel (b) - the requirements for acceptability can also be framed in terms of statistical power. Regardless of δ, under the standard alpha ($\alpha = .05$), statistical power needs to be above .92 for CIs over δ to exclude ψ values less than 3. Minimally acceptable estimation of an effect requires its detection to be near certain: Common but arbitrary power standards, such as .80, do not yield estimates that rule out unacceptable estimation accuracy.
- Panel (c) - it is well known that a larger n results in smaller observed effects becoming statistically significant. However, the ψ associated with said effects can still be unacceptable. For example, critical effects with a p of .05 yield a ψ of approximately 5, with CIs that include values very close to 1. In comparison, critical effects with a p of .0005, which are approximately 78% larger than their .05 counterparts, yield confidently acceptable ψ values. We note that using an α of .0005 as a threshold for null hypothesis testing is a stricter standard than other recent proposals that focus on the detection of effects (Benjamin et al., 2018). Such a stringent criterion makes it more difficult for questionable researcher practices, such as

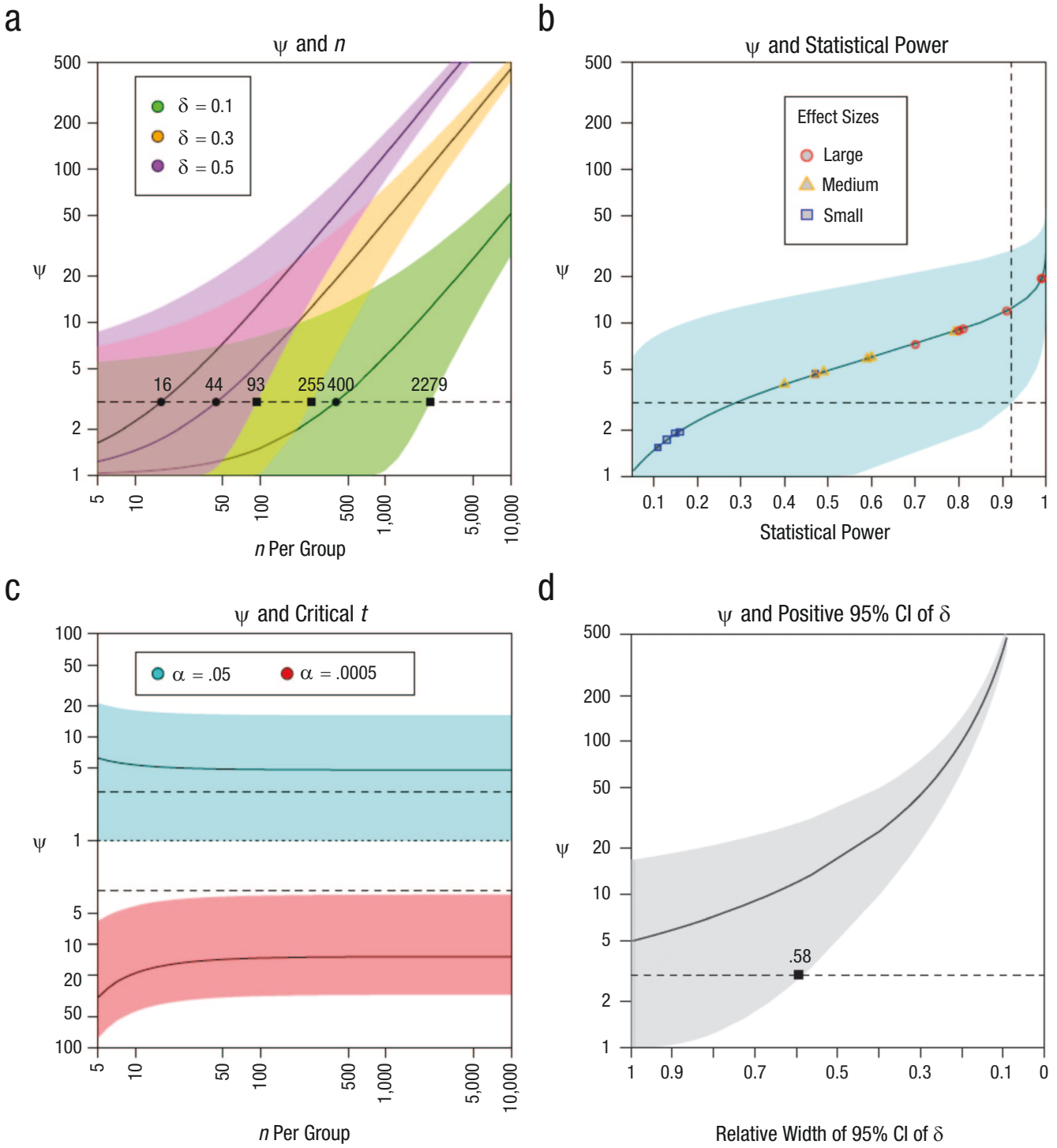

Fig. 2. Relationship between ψ and different relevant quantities. The bands correspond to the 95% confidence intervals (CIs) of ψ. The power values reported in (b) are also reported in Table 1. For further details, see the main text and the Appendix.

p-hacking (Simmons et al., 2011), to affect the outcome. Finally, these results may also serve to dampen researcher urges to characterize nonsignificant effects ($p > .05$) as if they are acceptably accurate.

- Panel (d) - some researchers consider an effect to be robust or reliable when the 95% CI of δ does not cross zero (Cumming, 2013). But when we transform a strictly positive or negative interval onto a range of plausible ψ values, we see that

Table 1. Median Power to Detect Small ($\delta = .2$), Medium ($\delta = .5$), and Large ($\delta = .8$) Effects as Reported in Meta-Analyses and Their Corresponding ψ Values (in Brackets).

	Small effect	Medium effect	Large effect
Meta-analyses	Median power [ψ]	Median power [ψ]	Median power [ψ]
Szucs & Ioannidis (2017)			
Cognitive neuroscience	0.11 [1.54]	0.40 [4.06]	0.70 [7.56]
Psychology	0.16 [1.94]	0.60 [6.13]	0.81 [9.48]
Medicine	0.15 [1.86]	0.59 [6.00]	0.80 [9.32]
Nuijten et al. (2020)			
Intelligence	0.11 [1.54]	0.47 [4.75]	0.99 [19.88]
Gaeta & Brydges (2020)			
Speech and language	0.13 [1.70]	0.49 [4.86]	0.91 [12.52]
Siegel et al. (2021)			
Industrial and organizational psychology	0.47 [4.58]	0.79 [8.86]	0.99 [19.88]

Note: The ψ values are also illustrated in Figure 2b. We calculated power for Gaeta & Brydges (2020) and Siegel et al. (2021) on the basis of median sample sizes.

they will include unacceptable values (for a threshold of 3) unless the width is less than 1.16δ (i.e., 58% of its maximum width of 2δ). In short, estimation accuracy can be unacceptable even for robust or reliable effects.

Examining Prior Meta-Analyses

We examined several recent meta-analyses to get a snapshot of how common poor ψ values are in various subfields (Gaeta & Brydges, 2020; Nuijten et al., 2020; Siegel et al., 2021; Szucs & Ioannidis, 2017). Table 1 shows a remarkable consistency across subfields, with the estimated median power to detect a small effect ($\delta = 0.20$) ranging between 0.11 and 0.16. These power estimates translate to ψ values ranging from 1.54 to 1.94, which strongly resemble the unacceptable scenario illustrated in the bottom row of Figure 1. Said simply, the majority of studies examining small effects in these fields may be producing results that are virtually indistinguishable from random conclusions. These meta-analytic values are also plotted in Figure 2 (b), where we show that even representative studies examining medium and large effects are not sufficiently powered to rule out unacceptable estimation accuracy.

Extensions

Our Lab 1 and Lab 2 framing provides a concrete way for scientists to grapple with inherently difficult questions about acceptable estimation accuracy and replication within the behavioral sciences. This framing could be extended to other estimators, testing frameworks, and experimental designs. In the current application, we focused on sample means and the usage of the independent two-sample *t* test. We did so because of the ubiquity of this experimental design and testing framework within the behavioral sciences. Our RCE formulation could be used to calibrate power and hypothesis-testing thresholds for statistical tests other than the standard *t* test, such as Welch's test, which allows for differences in group variance (Welch, 1947). Future work could explore how different configurations of group variances impact the RCE sample-mean comparison and what testing and power thresholds provide acceptable estimation accuracy.

The RCE is defined by the randomization of group labels on the estimates of interest, but these are not required to be population means. In keeping with our two-group design, an RCE could be defined as the randomization of group labels to estimates of population medians, which may be an interesting application for heavily skewed distributions. One could then examine alternative power and hypothesis-testing calibrations for tests such as the Wilcoxon-Mann-Whitney *U* test. It should be noted, however, that the Wilcoxon-Mann-Whitney *U* test is appropriate only for evaluating whether two population medians are different under relatively strict assumptions—that is, that both populations are identically distributed and differ only by a shift in location (Divine et al., 2018).

The RCE and two-labs perspective could be extended to other experimental designs. In defining a general RCE comparison, we want to preserve two distinct features of our current formulation. First, a generalized RCE should randomize the conclusions of scientific

interest. Applications could include a one-way analysis of variance, in which group mean labels are randomized, thus randomizing which means are larger than others while preserving Type I and Type II error rates for the omnibus F test. Generalizations could also include multiple regression: Certain aspects of the estimation process could be randomized, such as whether one standardized regression coefficient is larger than, or has the same sign, as another.[4] Second, a generalized RCE should also yield statistically significant results at rates similar to the standard estimation method being evaluated. This gives a generalized Lab 2 comparison additional bite, because the generalized RCE is not just randomizing the direction of results; it is also leading to random decisions regarding data. This second point is not intended to avoid important questions relating to preregistration practices (Nosek et al., 2019; Szollosi et al., 2020) but rather to place a finer point on an RCE comparison.

Given a suitable RCE and a standard method of estimation (e.g., ordinary least squares), we define a generalized ψ as the ratio of the respective mean-squared-error values. Although MSE has several nice properties, other accuracy metrics could also be substituted. Under this definition, ψ retains its simple interpretation: An estimator is ψ times as accurate as a generalized RCE. Future work could develop these comparisons and relate them to existing techniques, such as CIs, statistical power, and hypothesis testing.

Recommendations

Report ψ intervals

When reporting CIs over Cohen's d values, we recommend also reporting the requisite ψ interval using that study's sample size. A CI communicates a range of plausible effect sizes, whereas the CI over ψ communicates how well the effect is being estimated relative to an easily understood benchmark. If the ψ CI includes values less than 3, it is worth reporting that the data do not rule out unacceptable levels of estimation accuracy. Although we have illustrated some consequences of using 3 as a threshold for ψ, other values could be used depending upon the context.[5] The key takeaway is that ψ intervals translate effect-size estimates into a comprehensible measure of estimation accuracy. Reporting ψ intervals also provides researchers a degree of nuance when reporting results, allowing them to claim (or not) the detection of an effect, up to the usual Type I error rate under a specified α level, while also being transparent about estimation accuracy. To be clear, no additional inference is taking place: Transforming a CI over δ values into one over ψ values is expressing the same information again from an estimation perspective. Making use of such a perspective can be done regardless of one's statistical-inferential inclinations (e.g., Bayesian vs. frequentist). It is worth noting once again that ψ is distribution-free, in that its interpretation as the ratio of MSE values between sample means and the RCE does not depend upon any particular distributional form (see the Appendix for details).

Power statistical tests for estimation

When conducting a priori power analyses, we recommend that the sample size be selected according to effective estimation of the effect, rather than simple detection. We demonstrated that power of .92, when using an α of .05, results in CIs over δ that exclude ψ values less than 3. This perspective offers a grounded rationale for power values, rather than the highly arbitrary, but quite common, value of .80. Selecting sample sizes in this way is similar in spirit to the work of Gelman and Carlin (2014) and connects to the work of Kelley and Maxwell (2003) and Kelley and Lai (2011), who argue for determining sample size on the basis of CI width. See also the work of S. F. Anderson et al. (2017), who present a power-analysis framework that incorporates publication bias.

Bayesian estimation

One takeaway from our arguments is that there simply is not much information contained in small samples and small effects. Bringing more information to the analysis can take many forms, with Bayesian methodology being an obvious approach. Informative priors can be used to improve estimation accuracy of mean estimates (Gelman et al., 1995), and such priors can be incorporated into the t test itself (see, notably, Rouder et al., 2009; Gronau et al., 2019; and Ly & Wagenmakers, 2021). Bayesian formulations are well suited for integrating informative hypotheses with cognitive models (Lee & Vanpaemel, 2018; Vanpaemel & Lee, 2012), which can help avoid some of the estimation issues we raise here. This approach is especially important for researchers who face limited sample sizes by the very nature of their investigations. Of course, the accuracy of Bayesian approaches under limited sample sizes will be prior dependent (e.g., McNeish, 2016). The Appendix also provides two examples of how prior beliefs can be incorporated into the computation of ψ.

Computational modeling and formal theory

Throughout, we have treated the accurate estimation of an effect as a primary goal. There is much to say about

whether conceptualizing and testing theories in this way is optimal from a meta-science perspective. Indeed, Scheel (2022) argued that many psychological hypotheses are imprecisely specified, leading to questionable attempts at replication and measurement. Improved theory and quantitative modeling can lead to more compelling tests (e.g., model selection; for a recent review, see Myung & Pitt, 2018), avoiding simple effect-based characterizations (van Rooij & Baggio, 2021); see also Guest and Martin (2021) and Proulx and Morey (2021). Lee et al. (2019) and Devezer et al. (2019) provide thoughtful analysis and argumentation for how formalism can be used to improve scientific practices.

A more stringent threshold (α = .0005) for two-group between-subjects hypothesis testing

Using $\alpha = .0005$ sets a more stringent threshold than recent high-profile recommendations for methods reform (Benjamin et al., 2018). It's hardly our goal to further contribute to file-drawer problems by arguing that some studies should not be published if ψ is less than 3. Indeed, we believe that all studies should be reported and that *p* values (likewise, ψ values) should not serve as gatekeepers to the literature. Yet for researchers who want to provide a characterization that goes beyond mere detection (e.g., "the two groups differ") and ensure that their estimates are distinguishable from random conclusions, a more prohibitive α level is arguably required. Rather than a tool for censorship, ψ can be perceived as a useful way to adjust the strength of one's claims to the expected accuracy of the estimation process.

The importance of experimental design

The fact that small effects are commonly observed does not mean that they are inevitable—one should always keep in mind the artificial and constructive nature of effects (e.g., Guala & Mittone, 2005; Woodward, 1989). In the behavioral sciences, effects are often small because of the use of minimal experimental manipulations that make the conditions being compared virtually identical, apart from a minor change (for a discussion, see Prentice & Miller, 1992). Researchers can rely on ψ to gauge the ability of a given experimental design to elicit a target phenomenon with sufficient accuracy, which in some cases can lead to the development of alternative experimental approaches. We do emphasize that notions of effect size are just one of many factors that impact experimental outcomes; see Buzbas et al. (2023) for a formal treatment of experimental design and its relation to replication rates.

Discussion

In reaction, one might argue that estimation accuracy should not be much of a concern if we care only about correctly detecting effects. We find this argument untenable for four reasons: First, knowledge about effect sizes plays a crucial role when using basic research findings to develop effective real-world interventions (Schober et al., 2018). Second, developing a theoretic account of the phenomena being studied typically requires more than just nominal or ordinal information (Meehl, 1978). Third, this reaction is at odds with the widespread use of statistical models that are predicated on quantitative comparisons of effects (Kellen et al., 2021), or the popularity of inferential frameworks that call for a quantitative reasoning of effects (Vanpaemel, 2010). Fourth, even in the context of coarse-grained theoretical accounts and ordinal predictions, knowledge about effect sizes is still relevant in the sense that it can inform us on matters of theoretical scope (i.e., how many people conform to a given theory's predictions; Davis-Stober & Regenwetter, 2019; Heck, 2021). That being said, we are not claiming that a focus on detection is by itself problematic, or that there are no legitimate contexts in which it takes center stage; we are asserting only that a mature scientific characterization calls for more than that, namely accurate estimates.

Alternatively, one could try to downplay the importance of estimation accuracy by arguing that talk of effects is by itself problematic, in the sense that effects are of secondary importance relative to the explanation of psychological capacities (van Rooij & Baggio, 2021). We take issue with pursuing such a line of reasoning here, as it mistakenly implies that giving psychological theorizing the attention that it is owed somehow eliminates effects from researchers' discourses. As a counterexample, consider the recent discussion on benchmark effects in short-term and working memory, a research domain that stands out for its highly sophisticated theoretical accounts (Oberauer et al., 2018). By contrast, the empirical exigencies of theory testing and development give estimation accuracy center stage (Meehl, 1978).

One could also argue that there is nothing new to see here, given that ψ is so closely related to already-established quantities. For instance, it is easy to see that ψ is a quadratic function of the *t* statistic (for details, see the Appendix). Rather than an all-new, all-different quantity to be reconciled with all the other ones in researchers' toolboxes, what ψ offers is a reframing of an old problem. It is an attractive feature, not a shortcoming,[6] that ψ is closely related to known quantities or tests, or that the pursuit of estimation accuracy ends up recovering similar methodological proposals with

distinct motivations (e.g., Benjamin et al., 2018). It is also worth noting once again that although we assumed Gaussian distributions when deriving our ψ value recommendations, the definition of the RCE and the subsequent interpretation of ψ as a ratio of MSE values is distribution-free.

Regardless of one's scientific view, random conclusions are indefensible. It follows that researchers' empirical findings should, at a minimum, be distinguishable from a foil whose conclusions are determined by a coin flip. But as we have demonstrated, this is easier said than done: Many published research studies, despite honest efforts, have barely improved upon the estimation accuracy of the infamous Lab 2. As it turns out, one can easily fail to reliably outperform Lab 2, even if effects are real, studies are based in strong theory, and no questionable research practices are at play. The RCE approach and the ψ index that can be derived from it provide a new perspective on methodological reform (Devezer et al., 2019; Munafò et al., 2017; Shrout & Rodgers, 2018). Everything begins with a simple statement: The estimation accuracy of our methods should be distinguishable from a random-conclusions foil. In the pursuit of this modest goal, we find that the default p value threshold of .05 does not rule out unacceptable conditions (see the bottom row of Fig. 1), leading us to more stringent criteria that also address known concerns with measurement error, statistical power, and replicability (Gelman & Carlin, 2014; Loken & Gelman, 2017; Maxwell et al., 2015; but see also Bak-Coleman et al., 2022). Based on these results, we believe that ψ and the RCE approach more generally constitute an important tool in improving psychological science.

Appendix

Formal characterization of ψ

All code is available on the Open Science Framework (OSF) at https://osf.io/2hza8/?view_only=f679d2211a314f469118e2fa27111fea.

Let $X_{1\mathrm{A}}, X_{2\mathrm{A}}, X_{3\mathrm{A}}, \ldots, X_{n\mathrm{A}}$ be n-many independent, identically distributed samples from a random variable, $\boldsymbol{X_{\mathbf{A}}}$, with mean μ_{A} and variance σ^2, where σ^2 is finite. Likewise, let $X_{1\mathrm{B}}, X_{2\mathrm{B}}, X_{3\mathrm{B}}, \ldots, X_{n\mathrm{B}}$ be n-many independent, identically distributed samples from a random variable, $\boldsymbol{X_{\mathbf{B}}}$, with mean μ_{B} and variance σ^2. We assume that $\boldsymbol{X_{\mathbf{A}}}$ and $\boldsymbol{X_{\mathbf{B}}}$ are independent of one another. We quantify accuracy via mean-squared error (MSE):

$$\mathrm{MSE} = \mathbb{E}[(\hat{\mu}_{\mathrm{A}} - \mu_{\mathrm{A}})^2 + (\hat{\mu}_{\mathrm{B}} - \mu_{\mathrm{B}})^2],$$

where "$\mathbb{E}[\cdot]$" is the expectation operator and $\hat{\mu}_{\mathrm{A}}$ and $\hat{\mu}_{\mathrm{B}}$ are, respectively, estimates for μ_{A} and μ_{B}.

Result 1. The ratio of MSE values between the random conclusion estimator (numerator) and sample means (denominator) is equal to

$$\psi(\delta, n) = \frac{n\delta^2 + 2}{2}.$$

Proof. We first calculate the MSE of the random conclusions estimator (RCE):

$$\begin{aligned}
\mathrm{MSE}_{\mathrm{RCE}} &= \frac{1}{2} E[(\bar{x}_{\mathrm{A}} - \mu_{\mathrm{A}})^2 + (\bar{x}_{\mathrm{B}} - \mu_{\mathrm{B}})^2] \\
&\quad + \frac{1}{2} E[(\bar{x}_{\mathrm{B}} - \mu_{\mathrm{A}})^2 + (\bar{x}_{\mathrm{A}} - \mu_{\mathrm{B}})^2] \\
&= \frac{1}{2} E[\bar{x}_{\mathrm{A}}^2 + \bar{x}_{\mathrm{B}}^2 - 2\bar{x}_{\mathrm{A}}\mu_{\mathrm{A}} - 2\bar{x}_{\mathrm{B}}\mu_{\mathrm{B}} + \mu_{\mathrm{A}}^2 + \mu_{\mathrm{B}}^2] \\
&\quad + \frac{1}{2} E[\bar{x}_{\mathrm{A}}^2 + \bar{x}_{\mathrm{B}}^2 - 2\bar{x}_{\mathrm{B}}\mu_{\mathrm{A}} - 2\bar{x}_{\mathrm{A}}\mu_{\mathrm{B}} + \mu_{\mathrm{A}}^2 + \mu_{\mathrm{B}}^2] \\
&= \frac{1}{2}\left(\frac{2\sigma^2}{n} + \mu_{\mathrm{A}}^2 + \mu_{\mathrm{B}}^2 - 2\mu_{\mathrm{A}}^2 - 2\mu_{\mathrm{B}}^2 + \mu_{\mathrm{A}}^2 + \mu_{\mathrm{B}}^2 \right) \\
&\quad + \frac{1}{2}\left(\frac{2\sigma^2}{n} + 2\mu_{\mathrm{A}}^2 + 2\mu_{\mathrm{B}}^2 - 4\mu_{\mathrm{B}}\mu_{\mathrm{A}} \right) \\
&= \frac{2\sigma^2}{n} + (\mu_{\mathrm{A}} - \mu_{\mathrm{B}})^2 \\
&= \frac{\sigma^2(n\delta^2 + 2)}{n}.
\end{aligned}$$

Equation 1 is obtained by taking the ratio of $\mathrm{MSE}_{\mathrm{RCE}}$ to the MSE of sample means,

$$\psi = \frac{\dfrac{\sigma^2(n\delta^2 + 2)}{n}}{\dfrac{2\sigma^2}{n}} = \frac{n\delta^2 + 2}{2}. \qquad \square$$

The value ψ is easily expressed in other metrics. It is equivalent to $t^2 + 1$, under the usual t metric, providing a direct relationship with the two-sample t test. Relevant to questions involving replication, we can also write out-of-sample R^2 (Campbell & Thompson, 2008), denoted R^2_{OS}, as a simple function of ψ. Consistent with typical formulations, we compare sample means against a competitor that uses the grand mean, $\overline{G} = \frac{1}{2}\overline{x}_{\mathrm{A}} + \frac{1}{2}\overline{x}_{\mathrm{B}}$, as the estimate for the population means in each group.

As before, we assume equal n in both groups. Direct calculation provides the following relationship:

$$R^2_{OS} = 1 - \frac{\text{MSE}_{\text{means}}}{\text{MSE}_{\bar{G}}} = 1 - \frac{\frac{2\sigma^2}{n}}{\frac{\sigma^2(2+n\delta^2)}{2n}} = 1 - \frac{4}{2+n\delta^2}$$

$$= 1 - \frac{2}{\psi},$$

where $\text{MSE}_{\text{means}}$ and $\text{MSE}_{\bar{G}}$ are the MSE values for sample means and the grand mean, respectively.

Comparing sample means to the RCE via Kolmogorov-Smirnov tests

We carried out a power analysis to determine the sample size n for achieving a power of .80 to reject the hypothesis that bivariate samples from the two distributions (sample means and RCE) are equal. We used the two-dimensional Kolmogorov-Smirnov test of Fasano and Franceschini (1987) with an α of .05. These power analyses were carried out in MATLAB using Lau's (2021) implementation of the test. The first row of Table A1 shows the required number of samples to achieve a statistical power of .80 as a function of ψ. We also carried out a power analysis using a one-dimensional test that examines the distribution of differences between mean estimates—that is, we calculated similar power analyses using $\bar{d} = \bar{x}_{\text{A}} - \bar{x}_{\text{B}}$. For this test, we used the two-sample Cramér-von Mises goodness-of-fit test (T. W. Anderson, 1962), as implemented in MATLAB by Cardelino (2021). The second row of Table A1 displays the required number of samples to achieve a power of .80 for each estimator as a function of ψ.

Rows 1 and 2 of Table A1 list the minimum number of studies (draws) per lab to reject the null hypothesis that estimates from the two labs follow the same generating distribution with a statistical power of .80. Row 3 presents the gain in information about the direction of an effect when estimated by sample means (relative to the RCE), where 0 represents no reduction in uncertainty and 1 represents total reduction in uncertainty.

Table A1. Power analyses.

	$\psi = 1.5$	$\psi = 2$	$\psi = 3$	$\psi = 5$
2D Kolmogorov-Smirnov test	54	31	19	15
Cramér-von Mises test	45	27	20	18
Information gain	.29	.49	.72	.91

Comparing samples means to the RCE via entropy

We can evaluate the two estimators with respect to information gain regarding the direction of the effect. The RCE randomly assigned condition labels according to a fair coin toss $(p_{\text{A}} = p_{\text{B}} = .50)$. The Shannon entropy of the RCE with respect to direction of the effect is given by

$$H(\text{RCE}) = -\{p_{\text{A}}\log_2(p_{\text{A}}) + p_{\text{B}}\log_2(p_{\text{B}})\}$$
$$= -0.50\log_2(0.50) - 0.50\log_2(0.50) = 1,$$

or total entropy about the direction. In other words, as n approaches ∞, RCE estimates converge to $\{-d, +d\}$. Note that entropy is not contingent on d or n and thus the RCE yields total entropy about the direction, regardless of sample size or effect size. The RCE thereby exemplifies the principle of maximum entropy (Jaynes, 1957), which holds that the probability distribution with the largest entropy best represents the most uniform state of knowledge. To that end, ψ contextualizes the sample-means estimator relative to an optimally deficient estimator, such that higher values of ψ indicate greater accuracy beyond mere random conclusions regarding direction.

The relationship between the two estimators can be further quantified in information-theoretic terms: As ψ increases from 1.0, the sample-means estimator will afford an increase in information relative to the RCE. This is illustrated in Figure A1, which depicts information gain (or reduction in entropy) as a function of ψ. The *y*-axis represents the bits of information that are gained when using sample means rather than the RCE, with values ranging from 0 (no information gain; i.e., sample means are as equally uninformative as the RCE) to 1 (complete information gain; i.e., sample means eliminate 100% of the uncertainty that comes with using the RCE). For example, the dotted line in the figure shows that the threshold $\psi = 3$ is associated with a 72% increase in information. If researchers desire a 90% gain in information beyond the RCE, they must achieve a ψ greater than 4.76; a 95% gain in information requires a ψ greater than 5.99. To create this figure, we used the *entropy* package in R (Hausser & Strimmer, 2009) to calculate the Shannon entropy H, in bits, of the sample means and RCE distributions generated by all combinations of $\delta \in \{0, .01, .02, .., .90\}$ and $n \in \{20, 22, 24, ..., 200\}$. We then found the reduction in entropy $H_{RCE} - H_{SM}$ (i.e., the information gain) at corresponding values of

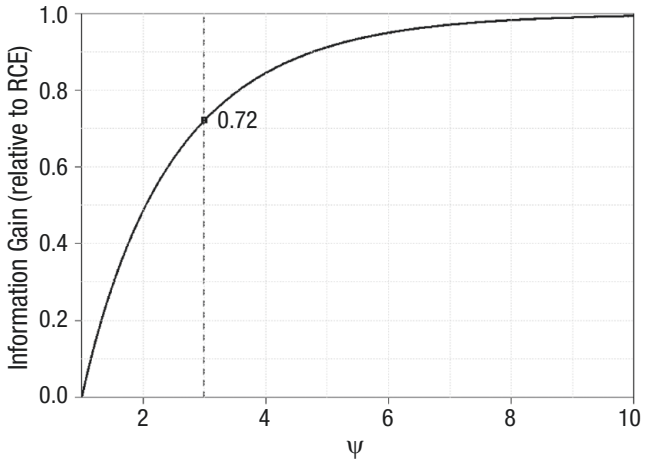

Fig. A1. Information gain afforded by sample means (relative to the RCE) regarding the direction of an effect as a function of ψ.

ψ. All code is available in the OSF repository linked above. Our use of information gain is equivalent to the (asymmetric) Kullback-Leibler divergence (Kullback & Leibler, 1951) $D_{\mathrm{KL}}(SM \parallel RCE) = \sum_{x \in \mathcal{X}} SM(x) \log\left(\frac{SM(x)}{RCE(x)}\right)$, which we deemed theoretically appropriate because it allows us to gauge improvements in the accuracy of sample means estimation relative to that of the maximally entropic RCE. Analysis of the (symmetric) Jensen-Shannon divergence reveals a nearly identical trajectory across values of $\boldsymbol{\psi}$, but without the theoretical alignment or ease of interpretation.

Quantifying the difference between distributions via the Wasserstein metric

To quantify the differences between the left and right sides of Figure 1 we relied on the Wasserstein metric, which is also known as the *Earth Mover's Distance* (EMD) because it determines the most efficient strategy for transporting a certain mass of earth from one position to another (Urbanek & Rubner, 2015). Specifically, the transportation of some mass from position $P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\}$, where p_i is a unit of the reference mass with weight w_{p_i}, to position $Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\}$, where q_j is a unit of the target mass with weight w_{q_j}, is given by the EMD: $\mathrm{EMD}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(d_{ij}f_{ij})}{\sum_{j=1}^{n} f_{ij}}$, where d_{ij} is the ground distance between p_i and q_j and f_{ij} is the optimal path from p_i to q_j.

In the current context, the EMD reflects the minimum amount of work (where one unit of work corresponds to transporting one unit of mass by one unit of distance) that is required to convert each random conclusions distribution to its corresponding sample-means distribution. We used the R package *emdist* (Urbanek & Rubner, 2015) to derive the EMD under each scenario in Figure 1 (see the OSF repository for code). When the effect size is small, the sample-means estimate in the bottom left panel (Fig. 1) shows that a negligible amount of work has been done to improve upon the random-conclusions estimate in the bottom right panel (Fig. 1), EMD = .106; on average, each unit of mass in the RCE panel would need to be moved just .106 units to match the mass in the sample-means panel. Relative to this small-effect-size condition, it would take six times more work to improve upon the RCE when δ is large (EMD = .635) and four times more work when δ is moderate (EMD = .424). In other words, more work is necessary whenever researchers want to ensure that their estimates are notably better than the mathematically least-informative estimate.

Confidence intervals around the true ψ

For each effect size δ considered, we computed 95% confidence intervals (CIs). The approach used to compute these intervals (see Cumming & Finch, 2001) consisted of determining the noncentral t distributions whose tails yield the observed t statistic (which can be obtained from δ and n) with nominal probabilities (e.g., 0.025 and 0.975). Because the present analysis focuses on absolute effect sizes, we established a lower boundary ($\delta = 0$) on these intervals (for which $\psi = 1$).

We investigated the coverage rates of the 95% intervals obtained for true effect sizes, such as the ones illustrated in Figure 2. Specifically, we performed the following steps:

1. Computed the 95% CI for known values of δ and n.
2. Generated $2 \times n$ samples from two normal distributions with variances 1 and means 0 and d.
3. Computed an effect-size estimate d from the $2 \times n$ samples taken in Step 2 and subsequently transformed this estimate into a $\hat{\psi}$ estimate.
4. Checked whether $\hat{\psi}$ was included in the CI computed in Step 1.
5. Repeated Steps 2 through 4 100,000 times.
6. Performed Steps 1 through 5 for different combinations of δ and n values.

The results reported in Table A2 show that the 95% CIs around the true effect sizes, when transformed into

Table A2. Coverage Rates of the 95% Confidence Intervals Around ψ for a Given True Effect Size δ and Sample Size per Group *n*.

	δ = 0.1	δ = 0.2	δ = 0.3	δ = 0.4	δ = 0.5
$n = 5$	0.93	0.94	0.94	0.94	0.94
$n = 10$	0.95	0.96	0.96	0.96	0.96
$n = 20$	0.96	0.97	0.97	0.97	0.97
$n = 50$	0.97	0.97	0.97	0.97	0.95
$n = 100$	0.97	0.97	0.96	0.95	0.95
$n = 200$	0.97	0.97	0.95	0.95	0.95
$n = 500$	0.97	0.95	0.95	0.95	0.95
$n = 1{,}000$	0.96	0.95	0.95	0.95	0.95
$n = 2{,}000$	0.95	0.95	0.95	0.95	0.95
$n = 5{,}000$	0.95	0.95	0.95	0.95	0.95
$n = 10{,}000$	0.95	0.95	0.95	0.95	0.95

ψ intervals, included the $\hat{\psi}$ estimates obtained from random samples roughly 95% of the time. These results corroborate our interpretation of these CIs around true values of ψ as ranges of plausible ψ estimates under a given effect size δ and sample size per group *n*.

Using effect-size priors to determine n

Further, ψ can be used to determine the sample size *n* that is expected to satisfy one's accuracy standards. Although our previous examples focused on point-effect-size values (see Fig. 2a), it is easy to incorporate prior beliefs in terms of an absolute effect-size distribution $\pi(\delta)$ with support over the positive reals. Let $\psi_{\min}$ be the minimum accuracy threshold. For a given effect size δ, the minimum *n* ensuring a threshold-satisfying ψ value is given by

$$n_{\min}(\delta, \psi_{\min}) = \frac{2\psi_{\min} - 2}{\delta^2}$$

The expected $n_{\min}$ can be obtained by calculating the following integral:

$$\mathbb{E}[n_{\min}] = \int n_{\min}(\delta, \psi_{\min})\pi(\delta)d\delta$$

For example, if we assume a uniform prior over $[0.1, 0.5]$ and a threshold of 3, then $\mathbb{E}[n_{\min}] = 80$. Note that, alternatively, one could consider the minimum *n* for a given δ that yields a lower bound of plausible ψ estimates that satisfies the threshold. If we consider the 95% CI as our range of plausible values, then an integration over δ like the one above yields $\mathbb{E}[n_{\min}] \approx 457$.

Finally, note that alternative prior distributions could be used instead (Gronau et al., 2019). For example, we could assume a truncated *t*-prior on δ with a location of 0.30, a scale of 0.05, degrees of freedom (*df*) of 3, and support over $[0.05, +\infty]$. This prior, which is illustrated in Figure A2, places most of its mass on effect sizes ranging between 0.2 and 0.4. Computing the above integrals using this informative prior instead results in $\mathbb{E}[n_{\min}]$ of approximately 56 and 323, respectively.

Histograms comparing the two distributions at different values of ψ

As an illustration, Figures A3–A7 display bivariate histograms for simulated data under sample means (left-hand columns) and random conclusions (right-hand

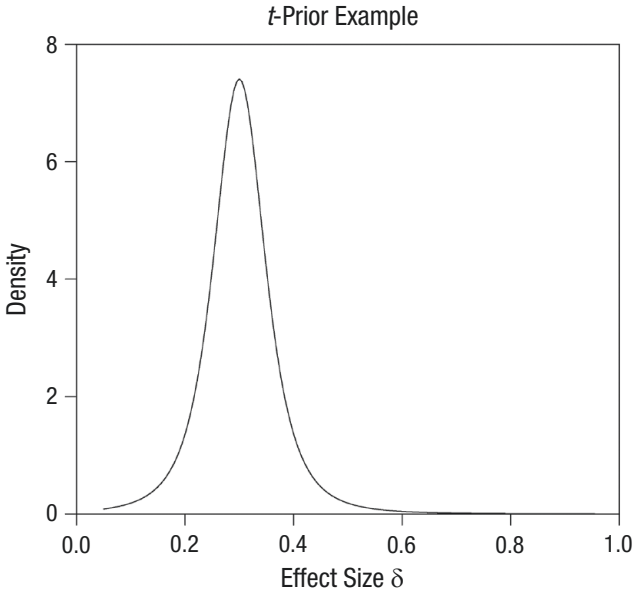

Fig. A2. Example of a truncated *t*-prior.

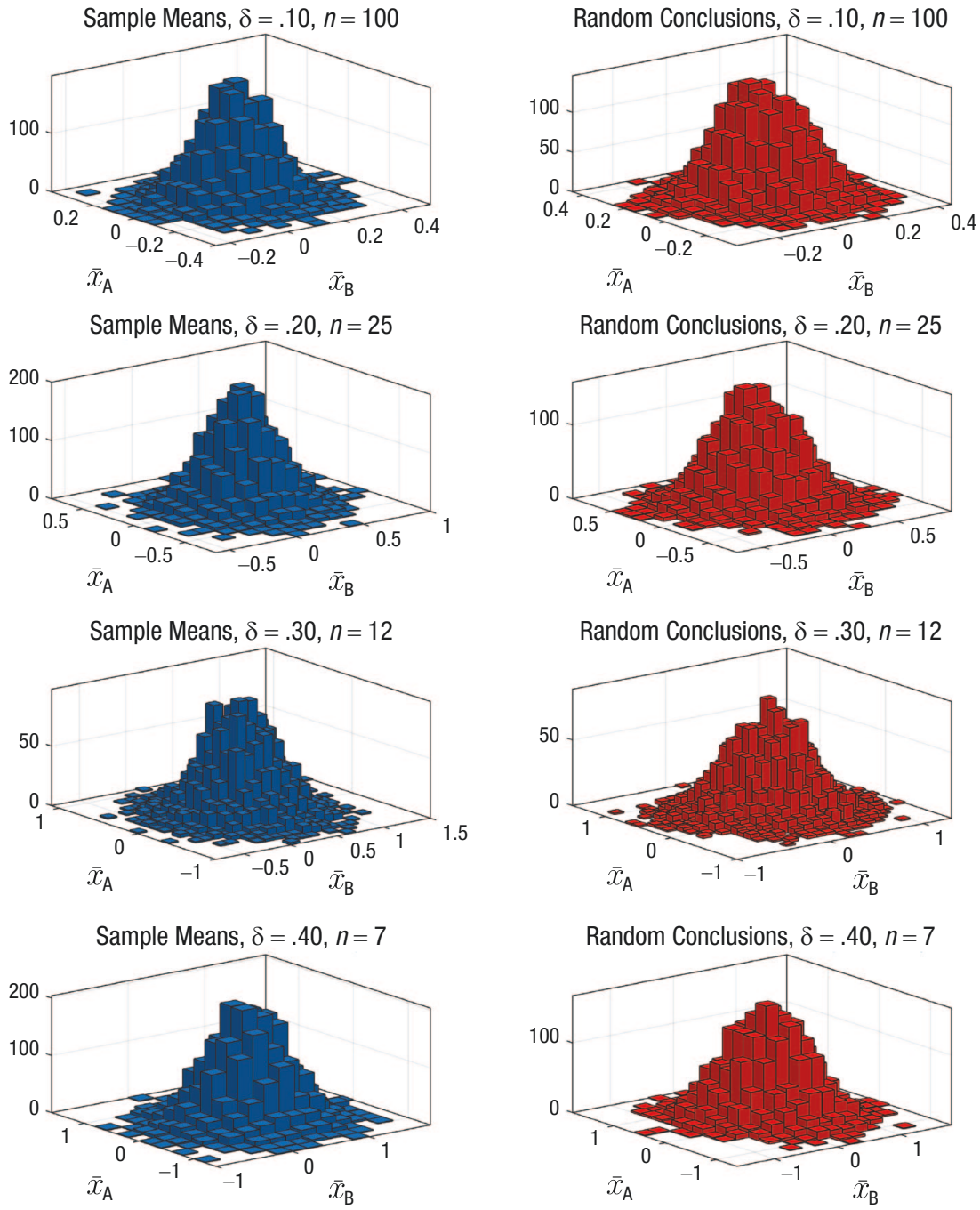

Fig. A3. Bivariate histograms comparing the sampling distribution of sample means to the random conclusions estimator under a ψ of 1.5. Each row of figures corresponds to a different combination of d and n to yield the same value of ψ.

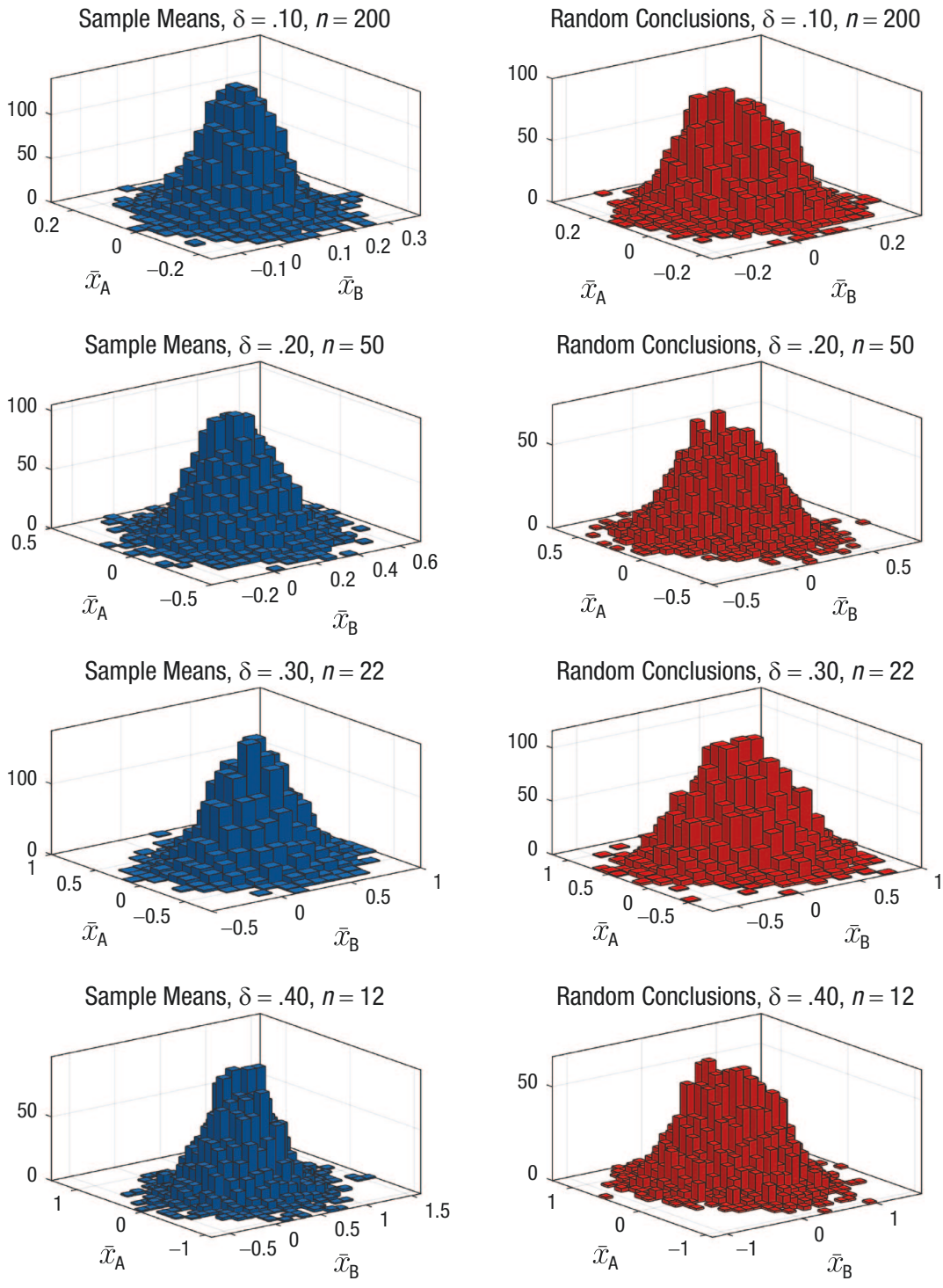

Fig. A4. Bivariate histograms comparing the sampling distribution of sample means to the random conclusions estimator under a ψ of 2. Each row of figures corresponds to a different combination of d and n to yield the same value of ψ.

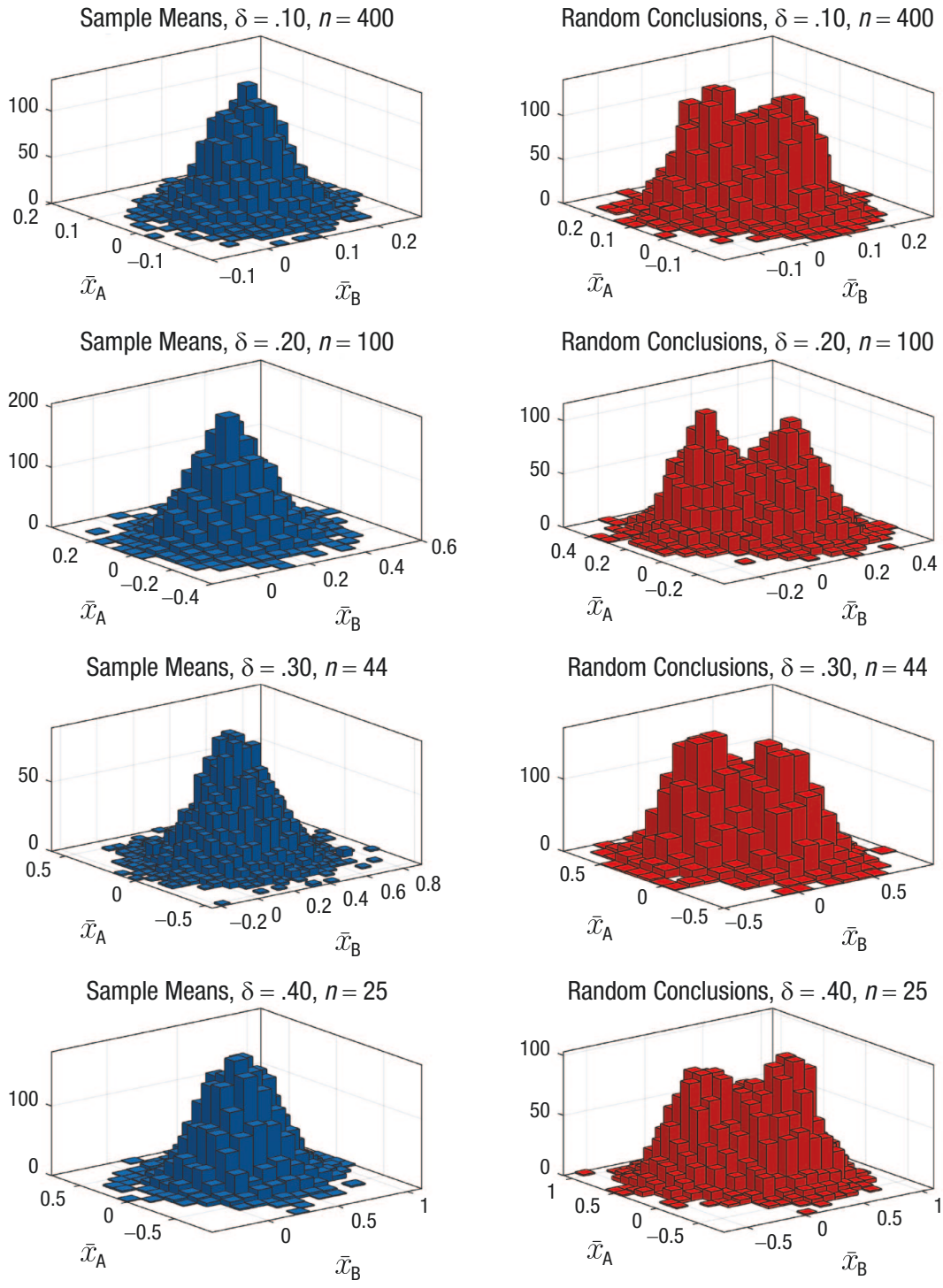

Fig. A5. Bivariate histograms comparing the sampling distribution of sample means to the random conclusions estimator under a ψ of 3. Each row of figures corresponds to a different combination of d and n to yield the same value of ψ.

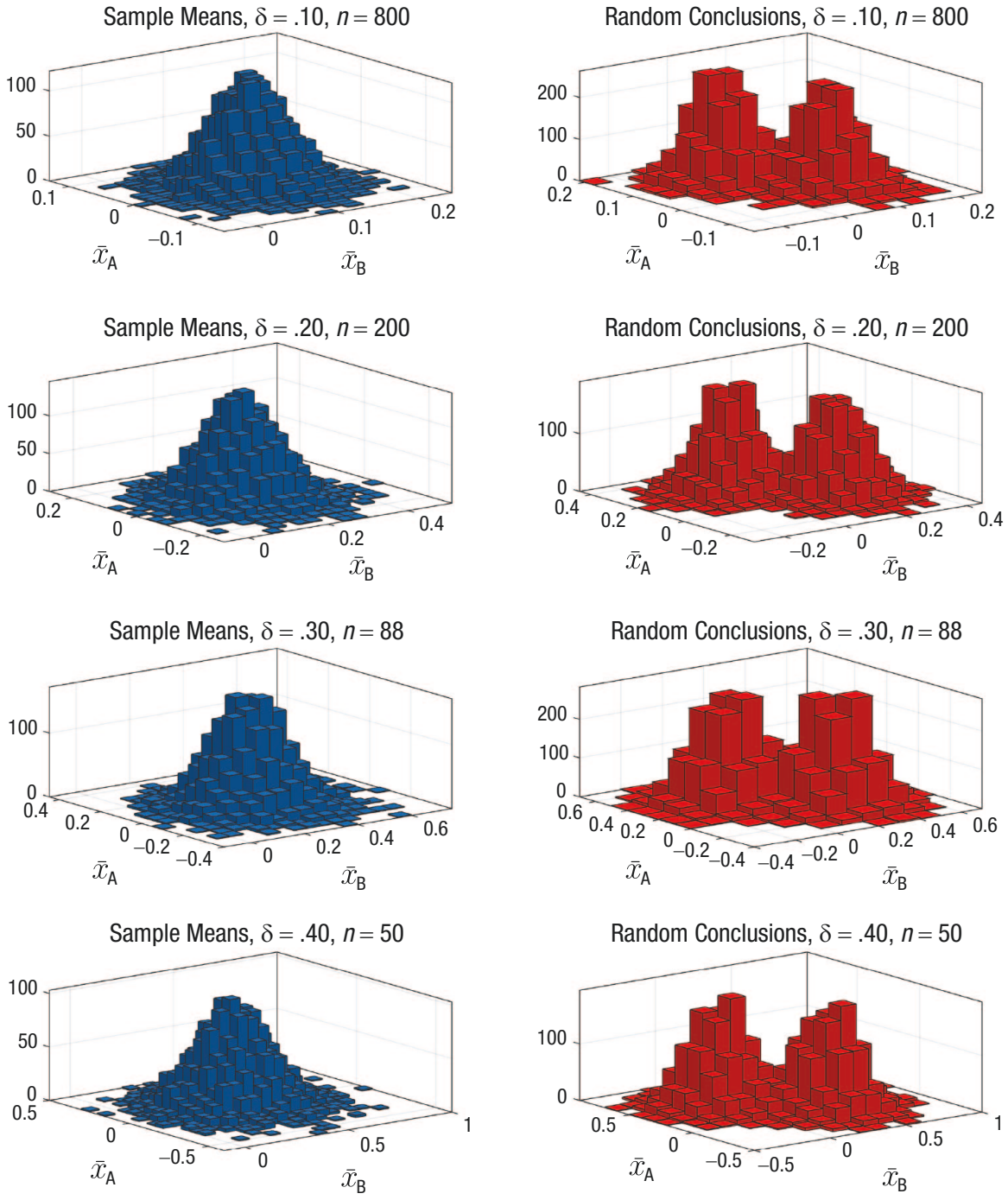

Fig. A6. Bivariate histograms comparing the sampling distribution of sample means to the random conclusions estimator under a ψ of 5. Each row of figures corresponds to a different combination of d and n to yield the same value of ψ.

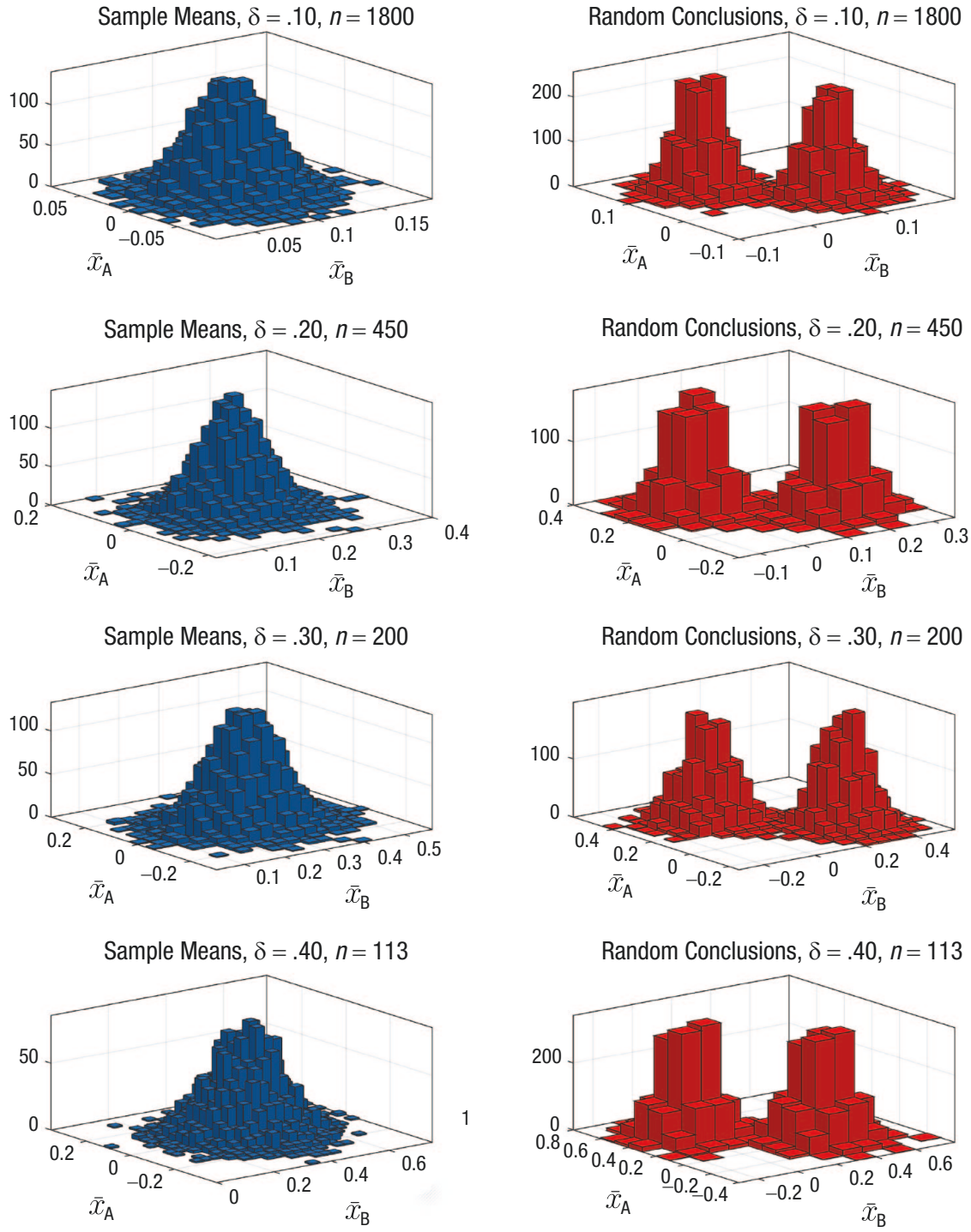

Fig. A7. Bivariate histograms comparing the sampling distribution of sample means to the random conclusions estimator under a ψ of 10. Each row of figures corresponds to a different combination of d and n to yield the same value of ψ.

columns) for ψ values of 1.5, 2, 3, 5, and 10. Each row corresponds to δ values of 0.10, 0.20, 0.30 and 0.40 for values of n that give the appropriate value of ψ. By examining Figures A3 through A7, we can see that the estimator's variance clearly depends upon n, but the relationship between sample means and the RCE remains stable for fixed values of ψ at different combinations of n and δ.

Transparency

Action Editor: Tim Pleskac
Editor: Interim Editorial Panel
Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

C. P. Davis-Stober acknowledges support from the National Institutes of Health under Grant No. K25AA024182 (NIAAA; C. P. Davis-Stober, primary investigator). W. Bonifay's contributions to this work were supported by the Institute of Education Sciences, U.S. Department of Education, through Grant No. R305D210032. D. Kellen's work was supported by a National Science Foundation (NSF) CAREER Award, ID No. 2145308. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Institutes of Health, Institute of Education Sciences, the National Science Foundation, the U.S. Department of Education, or the authors' home institutions.

ORCID iDs

Clintin P. Davis-Stober https://orcid.org/0000-0002-6836-6979
David Kellen https://orcid.org/0000-0003-1483-7875
Wes Bonifay https://orcid.org/0000-0003-3853-7607

Notes

1. This type of argument is foundational to myriad other existing procedures, such as determining whether fitted network models are distinguishable from networks with randomly determined connections (Steinley & Brusco, 2021). Looking further back, another example would be techniques such as Horn's parallel analysis (Horn, 1965).
2. We use δ to denote the true effect size in the population and d to denote sample estimates of δ. When we refer to the population value of Cohen's d, we are referring to δ.
3. To be clear, we are also not suggesting that comparisons with Lab 2 can serve as a way to identify errors or questionable research practices.
4. See Davis-Stober and Dana (2014) for a proto-RCE estimator along these lines.
5. Indeed, rejecting the null at the $\alpha = .05$ level is equivalent to stating that the 95% CI over ψ does not include 1, a value that can be achieved only if there is precisely no effect.
6. For a similar scenario in which the same model-selection index is derived from very different theoretical foundations, see Grünwald and Navarro (2009) and Karabatsos and Walker (2006).

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562.

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

Anderson, T. W. (1962). On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, *33*, 1148–1159.

Bak-Coleman, J., Mann, R. P., West, J., & Bergstrom, C. T. (2022). *Replication does not measure scientific productivity*. SocArXiv rkyf7, Center for Open Science. https://doi.org/10.31235/osf.io/rkyf7

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*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Buzbas, E. O., Devezer, B., & Baumgaertner, B. (2023). The logical structure of experiments lays the foundation for a theory of reproducibility. *Royal Society Open Science*, *10*, Article 221042. https://doi.org/10.1098/rsos.221042

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Cardelino, J. (2021). *Two sample Cramér-von Mises hypothesis test*. https://www.mathworks.com/matlabcentral/fileexchange/13407-two-sample-cramer-von-mises-hypothesis-test

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574.

Davis-Stober, C. P., & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: A new perspective on replication and the "crisis of confidence" in psychology. *Behavior Research Methods*, *46*, 1–14.

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox' of converging evidence. *Psychological Review*, *126*(6), 865–879.

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), Article e0216125. https://doi.org/10.1371/journal.pone.0216125

Divine, G. W., Norton, H. J., Barón, A. E., & Juarez-Colunga, E. (2018). The Wilcoxon–Mann–Whitney procedure fails as a test of medians. *The American Statistician*, *72*(3), 278–286.

Domingue, B., Rahal, C., Faul, J., Freese, J., Kanopka, K., Rigos, A., Stenhaug, B., & Tripathi, A. (2021). *Intermodel vigorish (IMV): A novel approach for quantifying predictive*

accuracy with binary outcomes. https://doi.org/10.31235/osf.io/gu3ap

Fasano, G., & Franceschini, A. (1987). A multidimensional version of the Kolmogorov–Smirnov test. *Monthly Notices of the Royal Astronomical Society*, *225*(1), 155–170.

Gaeta, L., & Brydges, C. R. (2020). An examination of effect sizes and statistical power in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1572–1580.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman & Hall/CRC.

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390.

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*, 205–215. https://doi.org/10.1177/1745691620984483

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, *74*(2), 137–143.

Grünwald, P., & Navarro, D. J. (2009). NML, Bayes and true distributions: A comment on Karabatsos and Walker (2006). *Journal of Mathematical Psychology*, *53*(1), 43–51.

Guala, F., & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, *12*(4), 495–515.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Hausser, J., & Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, *10*(7), 1469–1484.

Heck, D. W. (2021). Assessing the "paradox" of converging evidence by modeling the joint distribution of individual differences: Comment on Davis-Stober and Regenwetter (2019). *Psychological Review*, *128*(6), 1187–1196.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620.

Karabatsos, G., & Walker, S. G. (2006). On the normalized maximum likelihood and Bayesian decision theory. *Journal of Mathematical Psychology*, *50*(6), 517–520.

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science*, *16*(4), 767–778.

Kelley, K., & Lai, K. (2011). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research*, *46*(1), 1–32.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*(3), 305–321.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*(448), 1372–1381.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lau, B. (2021). *2-d Kolmorogov-Smirnov test, n-d energy test, Hotelling T^2 test*. https://github.com/brian-lau/multdist

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, *2*(3), 141–153.

Lee, M. D., & Vanpaemel, W. (2018). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, *25*(1), 114–127.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585.

Ly, A., & Wagenmakers, E.-J. (2021). *Bayes factors for peri-null hypotheses*. https://doi.org/10.48550/ARXIV.2102.07162

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., & Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*, 654–660.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 750–773.

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.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Myung, J. I., & Pitt, M. A. (2018). Model comparison in psychology. In J. T. Wixted, E. A. Phelps, & L. Davachi (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (Vol. 5, pp. 85–118). John Wiley & Sons, Inc.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28–34.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818.

Nuijten, M. B., van Assen, M. A., Augusteijn, H. E., Crompvoets, E. A., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. *Journal of Intelligence*, *8*(4), Article 36.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*, 885–958.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*(1), 160–164.

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, *16*(4), 671–681.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), Article 26.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development*, *31*(1), Article e2295.

Schober, P., Bossers, S. M., & Schwarte, L. A. (2018). Statistical significance versus clinical importance of observed effect sizes: What do *p* values and confidence intervals really represent? *Anesthesia and Analgesia*, *126*(3), 1068–1072.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.

Siegel, M., Eder, J. S. N., Wicherts, J. M., & Pietschnig, J. (2021). Times are changing, bias isn't: A meta-meta-analysis on publication bias detection practices, prevalence rates, and predictors in industrial/organizational psychology. *Journal of Applied Psychology*, *107*(11), 2013–2039.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Steinley, D., & Brusco, M. J. (2021). On fixed marginal distributions and psychometric network models. *Multivariate Behavioral Research*, *56*(2), 329–335.

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is preregistration worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95.

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

Urbanek, S., & Rubner, Y. (2015). *Emdist: Earth mover's distance*. R package version 0.3-2.

van de Schoot, R., Hoijtink, H., & Jan-Willem, R. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, *2*, Article 24. https://doi.org/10.3389/fpsyg.2011.00024

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W., & Lee, M. D. (2012). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic Bulletin & Review*, *19*(6), 1047–1056.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, *34*(1–2), 28–35.

Woodward, J. (1989). Data and phenomena. *Synthese*, *79*, 393–472.