aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2014, Vol. 9(3) 278–292

DOI: 10.1177/1745691614528520
pps.sagepub.com
SAGE

Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies

Daniël Lakens[1] and Ellen R. K. Evers[2]
[1]School of Innovation Sciences, Eindhoven University of Technology; and [2]Department of Social Psychology, Tilburg University

Abstract
Recent events have led psychologists to acknowledge that the inherent uncertainty encapsulated in an inductive science is amplified by problematic research practices. In this article, we provide a practical introduction to recently developed statistical tools that can be used to deal with these uncertainties when performing and evaluating research. In Part 1, we discuss the importance of accurate and stable effect size estimates as well as how to design studies to reach a corridor of stability around effect size estimates. In Part 2, we explain how, given uncertain effect size estimates, well-powered studies can be designed with sequential analyses. In Part 3, we (a) explain what *p* values convey about the likelihood that an effect is true, (b) illustrate how the *v* statistic can be used to evaluate the accuracy of individual studies, and (c) show how the evidential value of multiple studies can be examined with a *p*-curve analysis. We end by discussing the consequences of incorporating our recommendations in terms of a reduced quantity, but increased quality, of the research output. We hope that the practical recommendations discussed in this article will provide researchers with the tools to make important steps toward a psychological science that allows researchers to differentiate among all possible truths on the basis of their likelihood.

Keywords
induction, *p*-curve analysis, *v* statistic, confidence intervals, sequential analyses

One of the goals of psychological science is to differentiate among all possible truths on the basis of their likelihood. The most important way to achieve this goal in psychology is the empirical cycle, in which induction (on the basis of statistical inference) and deduction (on the basis of theory) take turns in progressive lines of research. Recent events have led psychologists to acknowledge that the inherent uncertainty encapsulated in induction is amplified by problematic research practices. Publication bias (Ioannidis, 2005) and flexibility during data analyses (Simmons, Nelson, & Simonsohn, 2011) create a situation in which false positives are easy to publish, whereas contradictory null findings do not reach scientific journals (but see Nosek & Lakens, in press). It is essentially impossible to predict whether a single statistically significant finding will replicate (Miller & Schwarz, 2011), and even when an effect is real, it is very common that the true effect size of a finding remains uncertain after initial studies have been published because of the large confidence intervals (CIs) surrounding effect size estimates (Cohen, 1994; Cumming, 2013). In this article, we provide a practical introduction to recently developed statistical tools that can be used to assess and mitigate these uncertainties when performing and evaluating research.

This article contains three parts. In Part 1, we discuss (a) how to report the uncertainty that surrounds effect size estimates with CIs, (b) how to plan studies such that a "corridor of stability" around an effect size estimate can be reached, and, more generally, (c) why collecting large enough sample sizes is necessary to reduce uncertainty. In Part 2, we focus on performing research given the inherent uncertainty about the true effect size of a hypothesized

Corresponding Author:
Daniël Lakens, Human Technology Interaction Group, School of Innovation Sciences, Eindhoven University of Technology, Room IPO 1.33, P.O. Box 513, 5600 MB Eindhoven, the Netherlands
E-mail: D.Lakens@tue.nl

effect, and we discuss how to run well-powered studies with sequential analyses and adaptive designs. In Part 3, we illustrate how researchers can evaluate past research by looking further than the statistical significance of the original findings. We discuss how individual studies can be evaluated on the basis of the accuracy with which they estimate effects in the population using the *v* statistic (Davis-Stober & Dana, 2013). Furthermore, we review how to correctly interpret a *p* value, what a *p* value conveys about the likelihood that an effect is true, how to use this knowledge to design studies in ways that increase the informational value of a *p* value, and how the evidential value of multiple studies can be examined with a *p*-curve analysis (Simonsohn, Nelson, & Simmons, in press). We hope these procedures (and the step-by-step guides in the Supplemental Material available online) will help researchers (a) evaluate the extent to which previous findings are worthwhile to cite or build on, (b) plan better studies, and (c) make an informed choice about whether to start a line of research with a close replication study (instead of a conceptual replication or extension) or whether previous research provides sufficient certainty about the likelihood of the hypothesis that a researcher can start with a conceptual replication.

Part 1: Confidence, Stability, and Reduction in Uncertainty

The certainty (or precision) of an effect size estimate increases with greater sample sizes. This precision is expressed by the variance of an effect size estimate. In general, the larger the sample size, the lower the variance (and thus the lower the standard errors) around estimates, such as means and effect sizes (for a detailed discussion, see Borenstein, Hedges, Higgins, & Rothstein, 2009). The variance of the effect size estimate is important because it largely determines the width of a CI. There is a direct relationship between the CI of an effect size and the statistical significance of the effect. For example, if an effect is statistically significant in a two-sided *t* test with an alpha of .05, the 95% CI for the mean difference between two groups will never include zero. CIs do not only provide information about the statistical significance of an effect but they also communicate the precision of the effect size estimate. Therefore, reporting CIs should not be considered as optional. As Kelly and Rausch (2006, p. 365) noted, "Although reporting measures of effect is useful, reporting point estimates without CIs to illustrate the uncertainty of the estimate can be misleading and cannot be condoned."

Although CIs communicate more information than a *p* value, the information they express is at least as (and perhaps more) difficult to understand. CIs express the proportion of intervals that, in the long run, will include the parameter for which the CI is calculated (which could be a mean difference, an effect size, or any other estimated parameter in a sample). In other words, in the long run, 95% of the CIs from a large number of close replications will capture the true population parameter. However, researchers typically have access to only a single 95% CI. The width of a single interval gives information about the precision with which a parameter estimate was measured. It also gives some indication about the likelihood that a similar value will be observed in future studies; however, when CIs are used for this purpose, it is risky. Simulations indicate approximately 83.4% (or five out of six) of replication studies will give a value that falls within the 95% CI of a single study;[1] although given publication bias or flexibility during the data analysis, this percentage is likely much lower for a randomly chosen 95% CI in the literature (Cumming, 2013).[2]

Calculating CIs for effect sizes requires an iterative procedure, which ESCI (Cumming, 2012) takes care of for Cohen's *d* and which can be done in SPSS for η^2 (or r^2) with scripts provided by Smithson (2001). In addition to the sample size, the width of the CI is influenced by precision of the measurement (measures that are inherently less accurate increase variance and, thereby, lead to less precise estimates of the effect size) and by the experimental design (in psychological experiments, within-subject designs typically provide more precision than between-subjects designs with the same sample size). We provide a practical primer and easy-to-use spreadsheets to calculate and convert between effect sizes *d* and *r* in the Supplemental Material (also see Lakens, 2013).

The corridor of stability

With a small number of observations, effect size estimates have very wide CIs and are relatively unstable. An effect size estimate observed after collecting 20 observations can change dramatically if an additional 20 observations are added. An important question when designing an experiment is how many observations are needed to observe relatively stable effect size estimates, such that the effect size estimate will not change considerably when more participants are collected. On the basis of approaches in statistics that stress accuracy, and not just statistical significance (e.g., Kelley & Maxwell, 2003), Schönbrodt and Perugini (2013) have recently performed simulations that address this question.

First, it is necessary to define the "stability" of an effect size estimate, and Schönbrodt and Perugini (2013) proposed that a useful (albeit arbitrary) benchmark is that the difference between the true and observed correlations should not exceed a small effect size as defined by Cohen (1988), neither in the collected sample nor in potentially increased samples. Hence, the stability of effect size estimates refers to a situation in which the

Table 1. Recommended Sample Size per Condition When Comparing Two Independent Groups Based for Different Effect Sizes (r and Cohen's d_{pop}) to Achieve the Point of Stability (POS) With 80% Confidence and Corridor Widths of .2 and .1 (see Part 1), to Achieve 80% or 90% Power to Observe the Effect With an Alpha of .05, and to Achieve a v Statistic Higher Than .5 (see Part 3)

r	d_{pop}	POS, 80% $w = .2$	POS, 80% $w = .1$	80% power	90% power	$v > .5$
.1	0.20	61	252	394	527	404
.2	0.41	57	238	95	126	99
.3	0.63	51	212	41	54	43
.4	0.87	43	181	22	29	23
.5	1.15	34	143	13	17	14
.6	1.50	25	104	9	11	9
.7	1.96	20	65	6	7	6

estimated effect is close to the true effect size and stays close. They defined the width (w) of the interval in which the effect size estimate has to stay using Cohen's q (the difference between two Fisher-z-transformed correlation coefficients), in which a q (or w) = .1 is defined as a small effect size. Schönbrodt and Perugini used the term *corridor of stability* to indicate effect size estimates that are likely to stay within a specified width around the true effect size, even when additional observations would be collected. The sample size associated with the point where effect size estimates enter the corridor of stability without leaving it (with a specific level of confidence) was referred to as the *point of stability*.

Large enough sample sizes

Because of the probabilistic nature of induction, statistical inferences require a large enough number of observations. With too small samples, researchers are (what we would call) "sailing the seas of chaos": Estimated effect sizes and significance levels are highly uncertain and can change dramatically as additional observations are added or from one study to the next (see also Stanley & Spence, 2014, this issue). In Appendix A, we provide a demonstration of the importance of large enough sample sizes, in which we randomly select subsamples of 50, 100, or 250 participants from a recent large-scale replication study (R. A. Klein et al., in press); we then show how the sample size determines (a) the variation in effect size estimates, (b) the percentage of studies that observe a significant effect, and (c) the overestimation in the effect size estimate if we only consider statistically significant studies.

To reach the corridor of stability, a researcher needs to collect a large enough sample size. On the basis of their simulations, Schönbrodt and Perugini (2013) provided a general recommendation of $n = 250$ per condition when examining effects of $r = .21$ (the average effect size in psychology according to Richard, Bond, and Stokes-Zoota, 2003) if researchers want to reach a small ($w = .1$) corridor of stability. We think it is also useful to consider the minimum number of participants per condition for effect sizes, ranging from $r = .1$ to $r = .7$, based on a wider corridor of $w = .2$. The choice of $w = .2$ is a lower bound based on the idea that for the average effect size in psychology ($r = .21$), an effect size estimate that is more than .2 lower will be in the opposite direction of the original hypothesis. In Table 1 (columns 1–4), we display sample size recommendations for a range of effect sizes to achieve points of stability with $w = .2$ or $w = .1$. Some readers might consider these sample sizes too large to be feasible, but this does not change the statistical reality: Large sample sizes are required to achieve stable effect size estimates.

Large enough sample sizes are also required to have high statistical power (the likelihood of observing a statistically significant effect if the effect truly exists). The sample sizes needed to achieve a statistical power of 80% and 90% for a two-sided independent t test with an alpha of .05, as a function of the size of the effect, are provided in Table 1. As is illustrated in Table 1, for large effect sizes, the goal to be accurate requires a larger sample size than the goal to have high power, whereas for small effect sizes, the goal to have high power requires larger sample sizes than the goal to be accurate (although, in general, smaller effect sizes require larger sample sizes).

We can use the sample sizes associated with points of stability to evaluate research findings. Large effects observed in small samples might not meet the minimum threshold for stability that we would like to see, which would be an indication that these effects should be interpreted with caution, and researchers interested in cumulative knowledge might want to start with a replication and extension study, instead of a conceptual replication, when building on such work. These sample sizes can also be used when designing new studies. New studies in which large effects are expected can be designed to reach the corridor of stability instead of reaching statistical significance (see also Maxwell, Kelley, & Rausch, 2008). In Table 1, we also point to the fact that if a researcher wants to perform a pilot study to acquire a reasonably accurate effect size estimate for an a priori power analysis, sample sizes per cell need to be quite large. A solution might be to perform an internal pilot study and to use sequential analyses, which we turn to next.

Part 2: Dealing With Uncertainty When Designing Empirical Studies

If the size of an effect is known, there is no need to study it. Even though the true effect size is inherently uncertain

when a study is designed, it is important to have at least some indication of the expected effect size to determine the sample size of a study. Sometimes, researchers use a heuristic to determine the planned sample size. A problem with using a heuristic is that the resulting sample size of a study is almost always suboptimal, in that either too many or, more likely, too few participants will be collected. A second solution is to use previous research findings in which a comparable effect has been examined and to use the effect size observed in these studies (or, preferably, the meta-analytic effect size of a large number of studies) to make a prediction about the hypothesized effect size. A problem with this solution is that, especially when meta-analyses are not available, effect sizes in the published literature typically have wide CIs and often overestimate the true effect size (see Lane & Dunlap, 1978). Even at their best, effect sizes observed in comparable studies are proxies of the expected effect size, which in reality might differ substantially because of the differences among studies.

A third solution is to perform a pilot study (the size of which can be based on the sample sizes associated with a point of stability with a $w = .2$; see Table 1) to estimate the effect size and to plan the sample size of the real study on the basis of the effect size observed in the pilot study. A problem with this approach is that studies with a small number of participants (which is typically the case in pilot studies) will provide relatively uncertain information about the true effect size. Therefore, power calculations based on the point estimate of an effect size in a pilot study will often be wrong. Researchers might also be tempted to combine data from the pilot study with that from the larger follow-up experiment (especially if the studies have identical methods). Unfortunately, not continuing with the follow-up experiment if there seems to be no effect in the pilot study can increase the Type II error rate, whereas not performing the follow-up study if the pilot study reveals a significant effect, but performing the follow-up experiment if the pilot study reveals no effect, can increases the Type I error rate (but see Sagarin, Ambler, & Lee, 2014, this issue, about post hoc corrections for unplanned interim analyses).

Sequential analyses

There is a different solution known as sequential analyses. The idea is straightforward. Instead of planning an experiment and analyzing the results when the data collection is complete, the data are analyzed intermittently. After an intermittent analysis, researchers follow preregistered decision paths based on the observed effect size. It is important to realize that repeatedly analyzing the data while data collection is in progress can increase the Type I error (see Simmons et al., 2011)—but only when the Type I error rate is not controlled. Statistical procedures to carefully control Type I error rates while performing sequential analyses have been developed in medical sciences and are widely used in large clinical trials. Psychologists can benefit from these techniques because they provide a practical and efficient way to deal with uncertain effect size estimates while guaranteeing well-powered studies. Imagine a researcher who expects to observe a small effect of Cohen's $d = 0.25$. An a priori power analysis (which can easily be calculated with the free G*Power software; Faul, Erdfelder, Buchner, & Lang, 2009) shows that an estimated 253 participants are needed in each between-subjects condition (or 506 participants in total) to have a statistical power of .80 with an alpha of .05 in a two-sided independent t test. Small variations in the observed effect size would have substantial consequences for the estimated sample size. With $d = 0.2$, the power analysis returns 788 participants in total, and with $d = 0.3$, only 352 participants are needed.

Sequential analyses provide a solution to this practical problem. When sequential analyses are used, researchers can spend the overall alpha level (typically .05) across the intermittent analyses. Different spending functions can be used depending on the goals of the researcher to determine the boundaries of the Z value and p value for each analysis. For instance, the O'Brien–Fleming rule is relatively conservative in the first analyses (when the effect size estimate is still sailing the seas of chaos), which is well suited for confirmatory tests, but a linear spending function is better suited for more exploratory research. When expecting an effect with $d = 0.25$, a conservative sample size plan would be to aim to collect 550 participants in total. Five sequential analyses could be planned after 110, 220, 330, 440, and 550 participants have been collected with the O'Brien–Fleming spending function. All the required computations can be performed with the freeware program WinDL, which was created by Reboussin, DeMets, Kim, and Lan (2000), or with the GroupSeq library in the free software R (step-by-step guides to perform all required calculations are included in the Supplementary Materials). This software provides us with the boundaries for our statistical tests at each of the five interim analyses, with upper bound Z scores of 4.88, 3.36, 2.68, 2.29, and 2.03, which correspond to p-value boundaries of .000001, .0007, .007, .022, and .042.

Imagine that after 330 participants, a t test returns the following: $t(328) = 2.74$, $p = .006$, Cohen's $d_s = 0.30$. Because the p value of .006 at the third analysis is smaller than the boundary ($p = .007$), the data collection can be terminated, and one can conclude that the data support the hypothesis. Although it is undesirable to design a study that has a statistical power of .5 at the final analysis, having a 50% chance to terminate the data collection

early after observing a significant result (or being able to continue the data collection when the test is not significant) will often be desirable. Sequential analyses will, on average, substantially save resources and can easily reduce sample sizes when examining true effects by at least 20% (see Lakens, in press).

Note, however, that the main goal of sequential analyses is to test hypotheses by demonstrating a statistically significant effect. Whenever researchers have the goal to provide accurate effect size estimates, they need to either design larger studies or turn to meta-analytic techniques. Moreover, stopping the data analysis early, if there is a significant difference, can lead to an overestimated effect size estimate; therefore, monitoring-adjusted p values, CIs, and effect size estimates are reported that have been developed to correct for this bias (see the Supplemental Material).

Sequential analyses are efficient and provide more flexibility. If researchers have specified a smallest effect size of interest, they can decide in advance to terminate the data collection whenever the observed effect size in an interim analysis is smaller than this value. In applied research, the smallest effect size of interest can often be determined on the basis of a cost–benefit analysis. For purely theoretical research, a researcher might have a computational model of the process under examination that suggests that an effect size lies between specific values. Alternatively, any effect size larger than zero might be considered of theoretical interest; however, in these situations, a researcher will often face practical limitations because of the number of participants that a researcher is willing to collect. In such situations, the smallest effect size of interest is determined by the minimum power (which is typically .8; Cohen, 1988) that can be achieved by collecting the maximum sample size that a researcher is willing to collect. For example, if a researcher decides that he or she is willing to collect no more than 200 participants to examine a hypothesis, this decision means that in practice the effect size that he or she could detect with 80% power in an independent t test would be a Cohen's d of 0.4 or larger. Especially when it is unclear whether an effect exists (e.g., for completely novel hypotheses), sequential analyses allow researchers to look at the data while data collection is in progress and to stop the data collection when the effect size is smaller than the smallest effect size of interest.

Researchers should be aware that terminating the data collection early when the effect is unlikely to exist always includes the risk of a Type II error. We recommend consulting the minimum sample sizes based on the corridor of stability (see Table 1) before taking effect size estimates with very wide CIs at an interim analysis too seriously (see Lakens, in press).

Adaptive designs

Sequential analyses also open up the possibility to use *adaptive designs*, in which the final sample size of a study is determined by a *conditional power analysis*. As opposed to an a priori power analysis, a conditional power analysis is performed on the basis of the effect size observed in the data that have been collected so far (which is referred to as an *internal pilot study*). By combining a smallest effect size of interest, sequential analyses, and an adaptive design, a researcher can preregister a study design in which he or she will (a) stop when the effect is significant at a predefined alpha level, (b) stop when the effect size reaches a value below the smallest effect size of interest, or (c) continue the data collection on the basis of a conditional power analysis. This flexibility is highly efficient. The only cost of using sequential analyses is a slight reduction in the power of the study, which should be compensated for by a small increase in the number of participants compared with a design without interim analyses (the exact increase depends on the number of interim analyses and the spending function that is chosen, which in this example leads to a 1% increase). For a more extensive introduction to sequential analyses, see Lakens (in press). For a detailed overview, see Proschan, Lan, and Wittes (2006).

Part 3: Evaluating Past Research Findings

When research findings are evaluated, the informational value of a study should always be taken into account, beyond the dichotomous conclusion of whether the statistical test performed on the reported data is significant. If an effect is observed reliably, it will most likely be statistically significant, but statistical significance does not necessarily imply that the effect is observed reliably. This asymmetry is important to understand, and it means that not every statistically significant effect is equally strong support for a theoretical line of reasoning. When meta-analyses are not available, researchers often need to judge the remaining uncertainty in a single or small number of published studies. There are several ways to evaluate published research findings; we highlight two. The first, the v statistic (Davis-Stober & Dana, 2013), can be used to interpret individual studies. The second, p-curve analysis (Simonsohn et al., in press), can reliably be used to interpret five or more studies.

Comparing estimates with guessing: The v statistic

Davis-Stober and Dana (2013) provided a new approach to evaluate the accuracy with which parameter values

(such as the average of a dependent variable) in the population are estimated on the basis of their corresponding observed values in a sample. They introduced the v statistic, which represents the likelihood of standard estimation methods to be more accurate than a recently developed benchmark that represents random guessing. The v statistic is the probability that a model based on the data is more accurate than a model in which the strength and direction of effects are randomly determined (for a conceptual primer, see Appendix B). The v statistic ranges from 0 (the model based on random guessing is always more accurate) to 1 (the model based on the data is always more accurate). Obviously, if a random estimator is more accurate than the estimator based on the observed data (indicated by a v statistic smaller than .5), a study does not really reduce the uncertainty about whether the hypothesis is true. It might seem that comparing empirical observations against random guessing is an extremely low benchmark, but as the v statistic demonstrates, the sample sizes used in some psychological studies can lead to the publication of findings that, although statistically significant, do not improve knowledge about the underlying structure of the data compared with random guessing (i.e., the v statistic is lower than .5). The v statistic is easily computed with scripts provided for the free software R (for a step-by-step guide, see the Supplementary Materials), is not defined in terms of null-hypothesis significance testing (NHST), and can be used to evaluate published research irrespective of unquantifiable doubts about researcher degrees of freedom or publication bias.

As an example of how the v statistic can be used, we turn to studies by Bargh, Chen, and Burrows (1996) and by Doyen, Klein, Pichon, and Cleeremans (2012), who examined whether people primed with words related to the elderly would walk more slowly through a corridor. We aim to show how the v statistic allows for a purely statistical evaluation of the performed studies. In Studies 2A and 2B, Bargh et al. reported a statistically significant difference in walking speed between the primed and control conditions, $t(28) = 2.86$, $p < .01$, and $t(28) = 2.16$, $p < .05$, respectively. The v statistic needs an estimate of the true effect size, and when it is calculated post hoc on the basis of observed values, r^2_{adj} (the unbiased estimator of the observed r^2) should be used.[3] An effect size estimate with the t value and total sample size in the studies by Bargh et al. can be computed that produces a Cohen's d_s of 1.04 and 0.79, or $r^2 = .226$ and .143, in Studies 1 and 2, respectively (assuming an equal distribution of participants across conditions). The v statistic for the total sample size (30 in each study), the number of estimated parameters (two in the case of an independent sample t test because two group means are estimated), and the computed effect size estimates ($r^2 = .226$ and .143) return a $v = .50$ and $v = .27$, respectively. These values imply that—given this sample size, estimate of the effect size, and number of parameters—the likelihood of the ordinary least squares estimator being more accurate than random guessing is 50% and 27%, respectively. In other words, with so few data points, any model based on the observed data lacks the accuracy that is needed for statistical inferences.

In a recent replication attempt, Doyen et al. (2012, Experiment 2) replicated the original effect; however, they only did so when experimenters expected the predicted effect to occur, which is in line with the possibility that the original effect was due to experimenter bias. Data from 25 participants revealed slower walking speeds in the primed condition ($M = 6.95$, $SD = 0.36$) compared with the nonprimed condition ($M = 6.52$, $SD = 0.31$), $F(1, 24) = 7.07$, $p = .014$, $r^2 = .228$. The estimation of these two means, which is based on data from 25 participants and yields an effect size of $r^2 = .228$, returns a $v = .44$. In other words, estimations based on the observed data in the study by Doyen et al. do not outperform an estimator that randomly determines the magnitude of the relationship between the means and is then scaled to the observed data. Elderly related primes might influence walking speed (but see Doyen et al., 2012, Experiment 1), and the explanation for the effect based on experimenter bias could be true; however, models based on the observed data are no more accurate than models based on random guessing. In other words, the data of both the original finding as well as the study in which the effect was observed only for experimenters who were aware of the hypothesis do not allow for an accurate model of the underlying effects; therefore, the data should be interpreted with caution until more convincing evidence has been provided.

A sufficiently high v statistic can be seen as a precondition before interpreting the outcome of a significance test. Researchers are already familiar with testing whether data meet certain conditions before statistical analyses are performed, such as whether data are normally distributed. Just as a p value is not very meaningful if the assumptions of the statistical test have been violated, a p value is not very meaningful when the v statistic is equal to or lower than .5. The v statistic thus functions as a useful benchmark to evaluate the accuracy of estimated means that are the basis for the statistical test. In contrast to post hoc power analysis, which is meaningless because it is directly related to the observed p value (Ellis, 2010), the v statistic is not directly related to the observed p value (i.e., studies can reveal a significant difference but still have a low v statistic, unlike the relationship between p_{rep} and p, or between p values and observed power) and can be used to retrospectively evaluate the appropriateness of the significance test.

The v statistic can be used as information about how strongly to weigh the statistical conclusions drawn from empirical observations. For example, if a researcher observes a statistically significant difference in a very small sample, the v statistic can help him or her to interpret whether this study was accurate enough to shift his or her belief about the likelihood that the effect is true or whether he or she should even interpret the outcome of the statistical test. Editors of scientific journals could use the v statistic as one of the factors that determine whether a study is worthwhile to publish. If a study is part of a line of research, it becomes possible to analyze a set of studies (which perhaps all have low v statistics) meta-analytically and to draw conclusions about the likelihood that effects are true on the basis of the cumulative data (see p-curve analyses, which are presented later in the article, or cumulative meta-analyses; Braver, Thoemmes, & Rosenthal, 2014, this issue). For between-subjects t tests, the minimum sample size in each condition to surpass a v statistic of .5 is provided in Table 1, and researchers should consider consulting this table when designing an experiment. The v statistic is relatively new, and its computations for within-subject designs have not yet been formalized, which would be a useful goal for future research.

Evaluating the evidential value of a single study

Statistical significance is a widely used criterion to decide whether an effect is interesting enough to share with other scholars. There has been much criticism of NHST (for a review, see Nickerson, 2000), but the use of NHST has not declined. Instead of arguing against the use of NHST entirely (e.g., Cumming, 2013), we briefly review how to correctly interpret p values, and we point out how they can be used when evaluating research findings.

In general, there are four possible outcomes of a statistical test. It is possible that a true effect is examined, in which case a significant result is a true positive, and a nonsignificant result is a false negative. If the effect under examination is not true, a significant result is a false positive, and a nonsignificant result is a true negative. The most prevalent misconception about p values is that they indicate the chance that an observed effect is a false positive (a mistake perpetuated through a number of introduction-to-statistics textbooks). Observing a significant effect with a $p = .03$ does not mean that the chance of a false positive is 3%, and it does not mean that it is 97% likely that the effect is true. In the following section, we review the information p values do provide, and we review the effect of statistical power and the prior likelihood that a hypothesis is true on the probability that significant results are Type I errors.

Because researchers sometimes investigate true ideas and sometimes investigate false ideas, it is important to understand what influences the likelihood that one of the four possible outcomes is observed. Increasing the statistical power of a study increases the likelihood of finding true positives while decreasing the likelihood of finding false negatives and, therefore, increases the informational value of studies. Assume that a researcher runs 200 studies in which novel research questions are examined and the alternative hypothesis (H1) and the null hypothesis (H0) are equally likely with the minimally recommended (Cohen, 1988) power of .8 (i.e., by examining a difference between two independent groups with a true effect size of Cohen's $d = 0.5$ with 51 participants in each condition). On average, he or she could expect to find 85 significant effects in these 200 studies—80 from 100 studies examining true effects, and five false positives from the 100 studies in which a nonexisting effect was examined (see Figure 1). Therefore, approximately 94% (80 out of 85) of the significant results are true positives, whereas 6% of the significant results are false positives. A nonsignificant result is also informative. Because only 20 of the 100 studies in which a true effect was examined did not yield a significant difference, and because 95 of the 100 studies in which a false idea was examined yielded nonsignificant results, only 17% (20 out of 115) of the nonsignificant findings are false negatives. If the researcher designs studies with a statistical power of only .35 (i.e., by examining a difference between two independent groups with a true effect size of Cohen's $d = 0.5$ with 21 participants in each condition), 12.5% of his or her significant results are false positives, and more than 40% of his or her nonsignificant results are false negatives. Running studies with low power thus decreases the informational value of significant, as well as nonsignificant, results.

Although it is common to implicitly assume that the null hypothesis and the alternative hypothesis are equally likely to be true (as we have done in the previous example), in many circumstances this assumption is not very reasonable. Some ideas are a priori more likely to be true than others. If one is testing highly unlikely (compared with likely) ideas, relatively more significant results will be false positives (for a detailed example, see Krzywinski & Altman, 2013). When designing studies that examine an a priori unlikely hypothesis, power is even more important: Studies need large sample sizes, and significant findings should be followed by close replications. Because only significant results end up in the published literature (due to publication bias), it is important to realize that surprising or unexpected findings from studies with small samples are more likely to be false positives than significant findings in which more likely hypotheses are examined (Ioannidis, 2005).

Fig. 1. Ratio of false-to-true positives and false-to-true negatives for the two researchers performing studies at 80% or 35% power. Both Researcher 1 and Researcher 2 conduct 200 experiments in which they examine 100 true ideas (in white) and 100 false ideas (in gray). Squares represent significant results, and circles represent nonsignificant results.

Evaluating the evidential value of a line of research

Even though a *p* value cannot be directly interpreted as the likelihood that an effect is true, this does not mean that *p* values do not provide any insight in the likelihood that H1 is true. There is a direct relationship between the power of studies and the distribution of *p* values from effects examined in these studies (see Cumming, 2008). Cumming (2008) has provided a formula that can be used to calculate these distributions in Excel.[4] When the null hypothesis is true (and the statistical power to observe a true effect is 0% because there is no true effect), *p* values are uniformly distributed. In other words, when there is no true effect in an experiment, every *p* value is expected to be observed with equal likelihood. This may seem counterintuitive, but a uniform distribution under the null is in essence the definition of a *p* value (for a more thorough explanation, see Cumming, 2008; Simonsohn et al., in press). When H1 is true, increasing the power of a study changes the distribution of *p* values from a uniform to a right-skewed distribution, resulting in relatively more small *p* values than large *p* values. Knowing the probability of observing a *p* value under H0 (when power = 0%) compared with the probability of observing that *p* value under the different distributions of power under H1 allows one to infer how much evidence for the H1 a specific *p* value provides.

For example, the probability of observing a *p* value between .00 and .01 is 1% under the H0, but when H1 is true in a study with 80% power, the probability has increased to approximately 59%. Assuming H0 and H1 are equally likely, a *p* value between .00 and .01 is 59 times more likely to be found under H1 (with 80% power) than under H0. The probability of observing a *p* value between .04 and .05 is also 1% under H0, but it is only slightly more probable under H1, in which it is 3% in a study with 80% power. Regardless of the power of a study, observed *p* values between .04 and .05 are never very likely, irrespective of whether H0 or H1 is true. Sellke, Bayarri, and Berger (2001) have calculated exactly how likely it is to find *p* values between .04 and .05. By looking at all possible distributions of *p* values and by calculating when the likelihood of finding a value

between .04 and .05 would be the highest, they found that the chance of finding such a p value is at best only 3.7%. Thus, whereas a p value between .00 and .01 is much more likely under H1 compared with H0, p values between .04 and .05 are at best only slightly more likely under H1. It is also interesting that when power is very high (approximately 96%), the probability of finding a p value between .04 and .05 is actually less likely under H1 than under H0 (because a considerably lower p value is much more likely). In other words, when power is very high, observing a (statistically significant) p value between .04 and .05 is actually more likely when there is no effect than when there is an effect. Even though researchers are trained to use p values as a dichotomous measure in which everything below $p = .05$ is called significant, researchers should actually weigh lower p values as stronger evidence for H1 than higher p values, especially in well-powered studies.

p-curve analysis

The distribution of p values depends on the power of the study. A single significant finding can be part of a distribution of other studies that revealed significant effects; however, because of publication bias, it can also be part of a set of mostly nonsignificant studies that researchers are not aware of. In the former situation, the effect is at least somewhat likely to be real, whereas in the latter case, the effect is more likely to be a false positive. Because the distribution of p values is different for true effects (i.e., right-skewed) than for null effects (i.e., uniformly distributed), it is possible to compare the distribution of an observed set of p values in the literature with the uniform distribution of a null effect and with the right-skewed distribution of a true effect, and it is possible to decide whether the observed distribution is more similar to that of a null effect or a true effect; p curve is a tool that does just that (Simonsohn et al., in press).

If the distribution of p values is significantly different from a uniform distribution and is right-skewed, the distribution of p values is consistent with a true effect. If the distribution is left-skewed, the studies have relatively (and significantly) more high p values (e.g., .04 and .05) than would be expected by chance if H0 is true (such a distribution is even more unlikely when H1 is true). If the distribution of p values does not significantly differ from the uniform distribution, the data do not provide evidence against H0. It is possible to examine whether the distribution of p values is flatter than expected if H1 is true (given a specific power of the performed studies). This allows researchers to test whether the effect is even smaller than a small effect. Simonsohn and et al. (in press) suggested testing whether the observed p-value distribution is flatter than would be expected if the performed studies were powered at 33% (but the test can be adjusted for different levels of power). If the distribution is significantly flatter than would be expected if studies had 33% power, the studies lack evidential value. The effect either does not exist or is much smaller than could be reliably observed in the performed studies.

Several characteristics of the p curve make it a very useful tool to estimate the evidential value of research. First, it is relatively easy to use. Second, any set of studies can be submitted to a p-curve analysis. Third, the power and false-positive rate of a p-curve analysis is very acceptable because p curve almost always has a higher power to correctly accept a set of studies as evidence for H1 than the individual studies. The probability with which p curve incorrectly rejects true effects depends on the size of the set of studies and their power, but by definition, the chance that a p-curve analysis incorrectly concludes that a set of studies has no evidential value is 5% when those studies were powered at the 33% level. Although it is unlikely that a p-curve analysis derives an incorrect conclusion from the available p values (see Simonsohn et al., in press), it is possible that the p-curve analysis remains inconclusive and that more p values are needed to determine whether a set of studies has evidential value.

Finally, and arguably most important, effects in the literature can be overestimations of the true effect sizes because of publication bias (e.g., Lane & Dunlap, 1978). Because a p-curve analysis is restricted to p values between .00 and .05, publication bias should not affect conclusions based on the distribution of p values within this range (though p hacking a true effect may cause the distribution to appear flatter than it should be). Indeed, p-curve analysis might even be a useful tool to obtain unbiased effect size estimates (Nelson, Simonsohn, & Simmons, 2014). This makes p-curve analyses a useful addition to a researcher's meta-analytic toolbox, complementing existing techniques such as meta-analyses (see Braver et al., 2014).

On www.p-curve.com, an easy-to-use interface and helpful instructions are available to select the right p values to perform a p-curve analysis. When planning to build on existing research, trying to consolidate inconsistent findings, or evaluating the likelihood of research findings for other reasons (e.g., as a reviewer or editor), p curve is a simple tool to accurately assess whether the reported studies contain evidential value for H1. When there is no evidential value for a theory, a researcher should proceed with caution when developing a line of research that builds on the analyzed ideas. If he or she decides to continue, lack of evidential value in a p-curve analysis suggests that the published literature might not provide the best inspiration for hypotheses. Instead, a better approach might be for the researcher to revisit the original theoretical ideas and to design the strongest test

Table 2. Guidelines to Increase and Evaluate the Informational Value of Studies

When planning studies:	Why?	How?	More information:
Power studies adequately . . .	Increases information by reducing the likelihood of false negatives and the relative likelihood of false positives in the published literature because of publication bias.	Always perform an a priori power analysis. When effect size estimates are extremely uncertain, use sequential analyses.	Cohen, 1988; Faul, Erdfelder, Buchner, & Lang, 2009 Lakens, in press; Proschan, Lan, & Wittes, 2006; see step-by-step guide in the Supplementary Materials
. . . aim for accurate effect size estimates.	Stable effect size estimate with narrow confidence intervals increase the confidence in the reliability of the effect.	Aim for the corridor of stability or a specific width of the confidence interval around the effect size estimate.	Maxwell, Kelley, & Rausch, 2008; Schönbrodt & Perugini, 2013
. . . aim for a v statistic > .5.	Make sure your model outperforms random guessing.	Collect enough participants, depending on the number of parameters and predicted effect size.	Davis-Stober & Dana, 2013
When interpreting studies:	**Why?**	**How?**	**More information:**
Look at the power of the study and the likelihood of the hypothesis.	Because of publication bias, low (estimated) power, and a low (estimated) likelihood of the hypothesis, the relative likelihood of false positives increases.	Never calculate post hoc power, but use an effect size estimate from a relevant meta-analysis or p-curve analysis. Make an informed (although subjective) judgment of the likelihood of the hypothesis on the basis of an evaluation of the theory.	Ioannidis, 2005; Nelson, Simonsohn, & Simmons, 2014
Look at the sample size.	Smaller samples yield less accurate and less reliable effect size estimates. With small samples, interpret results with caution.	Compare the sample size in the study with the recommended number of participants in Table 1.	Maxwell et al., 2008; Schönbrodt & Perugini, 2013
Is the model accurate?	The model should at the very least be better than random guessing.	Calculate the v statistic. The lower the v value, the less accurate the model is. When $v < .5$, interpreting the p value is not very meaningful.	Davis-Stober & Dana, 2013; see step-by-step guide in the Supplementary Materials
Is there evidential value in a set of studies?	Because of selection bias, the number of Type I errors in the literature might be larger than desired.	Look at the p curve of all results. Distribution of p values should be right skewed. A uniform (or left-skewed) distribution indicates a lack of evidential value.	Simonsohn, Nelson, & Simmons, 2014; see www.p-curve.com

of those ideas that he or she can come up with (perhaps using a completely different methodology). As an illustration of how p-curve analyses can differentiate between the evidential values of two ostensibly similar research lines, see Lakens (2014).

Discussion

Researchers are familiar with interpreting the outcome of empirical research based solely on the p value. We hope this article has explained why it is important to look at the sample size and the effect size of a study before looking at the p value. We have proposed the use of a number of relatively new statistical procedures and insights (see the checklist in Table 2), such as the corridor of stability (Schönbrodt & Perugini, 2013), sequential analyses (which are based on an old idea but on more recently developed statistical methods; e.g., Proschan et al., 2006), the v statistic (Davis-Stober & Dana, 2013), and p-curve analysis (Simonsohn et al., in press). It is often thought that editors and reviewers are conservative when it comes to new statistical techniques. In this article, we have

attempted to explain the immediate benefits of introducing new statistical procedures that allow researchers to evaluate and deal with the uncertainty that is inherent in an inductive science. The researchers who have created these novel techniques have provided easy-to-use guidelines and programs to incorporate these techniques correctly without much effort into one's workflow, and we provide easy to follow step-by-step guides in the Supplemental Material.

Especially when used in conjunction with methodological improvements—such as the preregistration of hypotheses, performing replication studies and meta-analyses, sharing findings that reveal nonsignificant differences, and publically posting data online (e.g., Asendorpf et al., 2013; Koole & Lakens, 2012; Nosek, Spies, & Motyl, 2012; Open Science Collaboration, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012)—researchers in our discipline can make important steps toward reducing the uncertainty about what is likely to be true. We believe the *v* statistic can be used to complement the use of the *p* value to interpret the informational value of studies, and *p*-curve analysis can complement more traditional meta-analytic procedures. Each of these techniques has its own strengths and weaknesses, and we recommend the use of multiple approaches to evaluate research. When designing studies, we provide recommendations for minimum sample sizes (see Table 1) if researchers have the goal to either observe a significant effect or the goal to provide an accurate effect size estimate, without presupposing that all individual published studies should meet either (or both) of these goals.

Coda

The practical recommendations discussed in this article have the goal to provide researchers with tools to deal with the uncertainty inherent in inductive sciences. We have to be honest about one important consequence for researchers who follow these recommendations: They do not give you the most bang for your buck when it comes to the number of significant findings you will observe in your studies. If your goal is to publish as many significant results as possible, it is more efficient to produce unreliable scientific knowledge than to produce reliable scientific knowledge (see also Bakker, van Dijk, & Wicherts, 2012; Nosek et al., 2012). This crucial insight should be clearly communicated to everyone who manages researchers on the basis of output measures, such as the number of publications.

Because running studies with large sample sizes is costly, and because the resources that a researcher has available are finite, a researcher is forced to make a trade-off between the number of studies that he or she runs and the power of these studies. The relation between sample size and power is a concave function, and therefore splitting a finite pool of participants over many low-powered studies will result in a higher total number of significant findings (even though the chance of finding a significant result is lower per study). This means that given the common practice of running underpowered studies (e.g., Button et al., 2013), increasing the quality of research necessarily decreases the quantity of publishable findings. If researchers are (or believe they will be) rewarded for the amount of significant findings that they produce, this system provides a perverse incentive to reduce the scientific quality of empirical studies and to increase the quantity of studies that will be much less informative about the truth.

A researcher running underpowered studies may have published more significant (but also hugely inaccurate) effects and may have contributed more Type I errors to the literature (especially when flexibility during the data analysis leads to an inflated alpha level). Most important, this researcher does not know which of the examined hypotheses were true but did not yield an effect because of low power, which studies yielded a significant result because the effect was actually true, and which significant results were actually Type I errors (see Figure 1). However, a researcher running fewer but larger studies may have fewer significant findings, but the effect sizes are estimated more accurately, and theoretical predictions that did not yield a significant effect have become less likely, because the studies had enough statistical power to observe an effect if there had been a true effect in the population. Furthermore, when an effect yielded a significant finding, it is more likely to be a true effect. All else being equal, a researcher running properly powered studies will clearly contribute more to cumulative science than a researcher running underpowered studies, and if researchers take their science seriously, it should be the former who is rewarded in tenure systems and reward procedures, not the latter.

There is no denying that a psychological science that considerably reduces the uncertainty about the likelihood that hypotheses are true requires larger sample sizes per study and will most likely reduce the number of novel empirical articles that a researcher can publish. As explained earlier, this outcome is actually desirable, given the goal of science to differentiate among all possible truths. It also makes other types of publications in addition to novel empirical work, such as meta-analyses or replications of important results (which can be accepted for publication before they are performed; see Nosek & Lakens, in press), more worthwhile, thereby providing further incentives toward a cumulative science. We believe reliable research should be facilitated above all else, and doing so clearly requires an immediate and irrevocable change from current evaluation practices in academia that mainly focus on quantity.

Appendix A

Variance in effect size estimates as a function of sample size

As an example of how more data provide a greater reduction in inductive uncertainty, we use data from a recent set of large-scale replication studies (R. A. Klein et al., in press), in which the *retrospective gambler's fallacy* was tested in one experiment (Oppenheimer & Monin, 2009). Participants were asked to imagine walking into a room and seeing a gambler roll three dice that either all come up 6, or roll two dice that come up 6 and one die that comes up 3. When estimating how many times the gambler had rolled the dice before the observed roll occurred, participants indicated a greater number of rolls in the condition in which three 6s were rolled. The replication study had 5,942 participants, and they also estimated more previous rolls in the condition in which three 6s were rolled ($M = 3.76$, $SD = 3.32$) than when two 6s and one 3 were rolled ($M = 2.07$, $SD = 2.07$), $F(1, 5940) = 576.36$, $p < .001$, with an effect size of $\eta^2 = .088$, 95% CI [.075, .102].

We randomly selected subsets of either 50, 100, or 250 participants from the total sample, repeating this selection 100 times for each sample size, and we tested the hypothesis in each subsample. The average results for the 300 samples are summarized in Table A1 as a function of the sample size. As the sample size increases, we move from the seas of chaos into the corridor of stability. The 100 effect size estimates are more similar in subsamples of 250 participants (as indicated by the lower standard deviation of the effect size estimate eta-squared) and vary more with smaller subsamples. Because of low statistical power, testing the hypothesis with 25 participants in each condition yields a significant difference in only 60% of the studies (with low power, p values have a wide distribution; see "the dance of the p values": Cumming, 2012). Although the average effect size estimates over 100 samples are comparable regardless of the size of the subsample, if we only calculate the average effect size for the statistically significant tests (mirroring a literature in which there is publication bias), then smaller sample sizes substantially overestimate the true effect size. As can be seen in Figure A1, the effect size estimates become more precise as sample sizes increase.

Table A1. Power, Percentage of Significant Results in 100 Randomly Selected Subsamples, Mean Effect Size in Subsamples and Statistically Significant Subsamples (Mirroring the Effect of Publication Bias), and the Standard Deviation of the Effect Sizes as a Function of the Size of the Subsamples

Subsample size	Power (%)	% $p < .05$	M η^2 (all studies)	M η^2 (significant studies)	SD η^2
50	58	61	.107	.153	.071
100	87	89	.096	.105	.048
250	99	100	.093	.093	.034

Appendix B

A brief conceptual primer on the v statistic

When we analyze data using a statistical procedure based on ordinary least squares (OLS) estimators (such as linear regression, analyses of variance, or any time we estimate population means by using the sample means), we are creating a model that fits our observed data best. Generally, models can have two types of error: error due to a bias in the estimate (the difference between the predicted mean and the observed mean) and error due to variance (the differences between different data points). OLS is an unbiased model, meaning it only has error due to variance. Constraining a model can decrease error in variance but can introduce error in bias. The v statistic introduces a benchmark of randomly generated (and thus arbitrary) constraints by randomly selecting both the relative strength of an effect as well as the direction. It then tests whether the error in this nonsensical model, the random least squares (RLS), is lower than that of a normal OLS estimate. Because such tests rely on the unknown true value of the population parameters, v is calculated for all possible parameter values that result in a probability of OLS being more accurate than RLS.

The v statistic can be seen as the probability of OLS being more accurate than a completely random (and therefore uninformative) model, and it ranges from 0 to 1. A v statistic lower than .5 means that the (random) RLS estimate is more likely to be an accurate estimate of the true effect than the OLS estimate. The error due to variance in any linear model (on the basis of OLS or RLS) is primarily influenced by the complexity of the model and the number of data points, whereas the error in bias in RLS is independent from the number of data points. Therefore, the more data, the more likely it is that OLS will outperform RLS. In other words, all that a study needs to outperform random guessing is a large enough sample size. For a more detailed and mathematical explanation of the v statistic, see Davis-Stober and Dana (2013).

Acknowledgments

We thank Geoff Cumming, Jason Dana, Clintin Davis-Stober, Alison Ledgerwood, and Felix Schönbrodt for helpful feedback on a draft of this article. We also thank numerous scholars for fruitful discussions on Twitter.

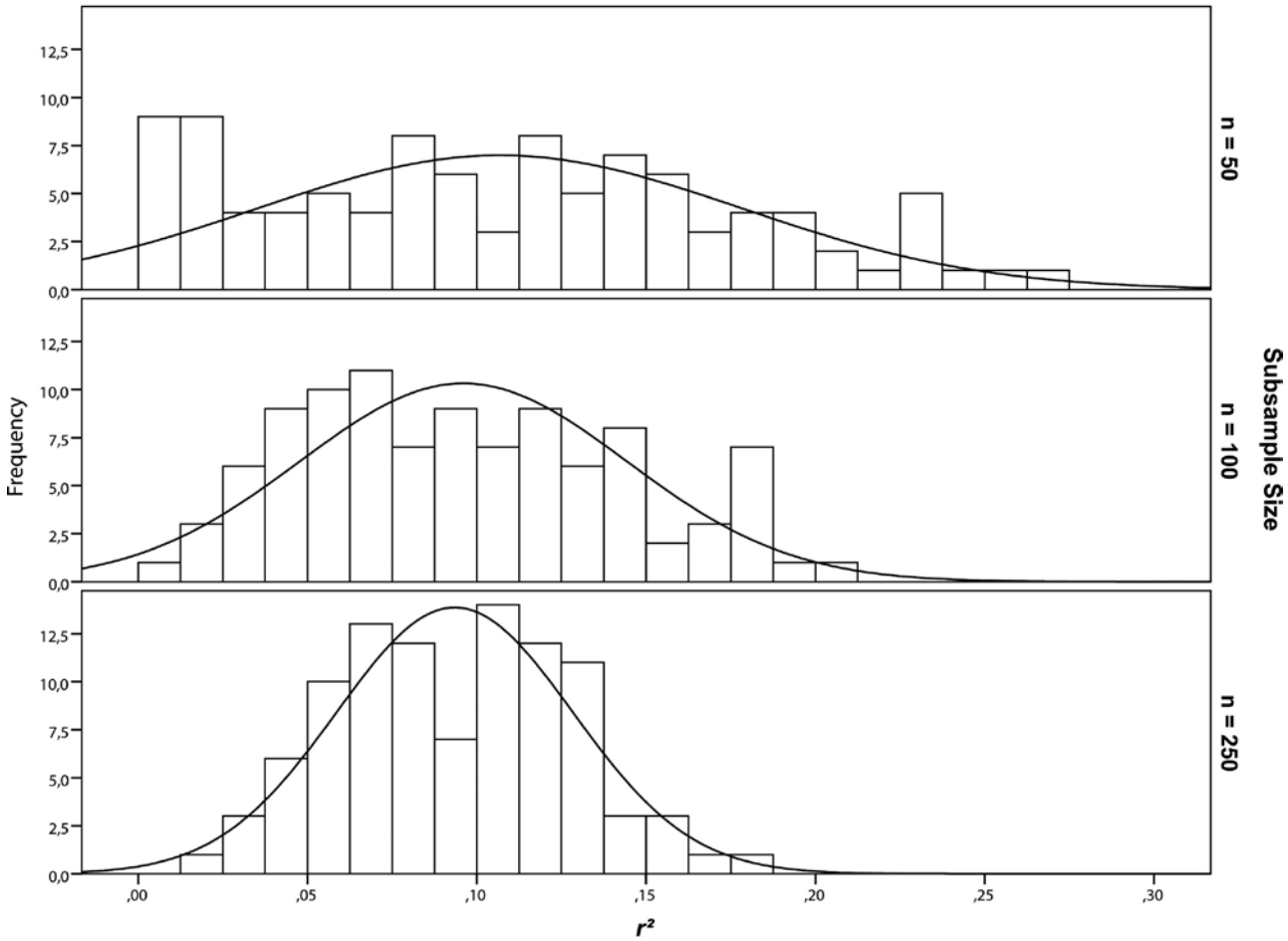

Fig. A1. Distribution and normal curves of 100 effect sizes from independent *t* tests performed on subsample sizes of 50, 100, and 250 participants, randomly drawn from the total data set by R. A. Klein et al. (in press).

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Supplemental Material

Additional supporting information may be found at http://pps.sagepub.com/content/by/supplemental-data

Notes

1. This percentage is 95% if, and only if, a parameter estimate is observed in a single experiment that is exactly the same as the true population value.

2. We should note that *p*-curve analysis (discussed later) is unaffected by publication bias and that the *v* statistic (also discussed later) is primarily determined by the sample size and, when calculated on the basis of an adjusted effect size estimate, is less affected by publication bias. Therefore, compared with CIs, *v* might be more useful to evaluate a single small study, and *p*-curve analysis might be the better approach to evaluate multiple studies.

3. The R script is available from http://psychology.missouri.edu/stoberc. See the step-by-step guide in the Supplementary Materials.

4. The formula can be found in Appendix B in Cumming's (2008) study. Note that in the formula to calculate z_μ on p. 299, a square-root sign is missing; the correct formula, $z_\mu = \delta\sqrt{(N/2)}$, is repeated on p. 300.

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230–244. doi:10.1037/0022-3514.71.2.230

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*, 333–342.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376. doi:10.1038/nrn3475

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G. (2013). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Davis-Stober, C. P., & Dana, J. (2013). 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. doi:10.3758/s13428-013-0342-1

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE*, *7*(1), e29081.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York City, NY: Cambridge University Press.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods*, *41*, 1149–1160. doi:10.3758/BRM.41.4.1149

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

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

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385.

Klein, R. A., Ratliff, K., Vianello, M., Adams, A. B., Jr., Bahník, S., Bernstein, N. B., . . . Nosek, B. A. (in press). Investigating variation in replicability: A "Many Labs" Replication Project. *Social Psychology*.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614. doi:10.1177/1745691612462586

Krzywinski, M., & Altman, N. (2013). Points of significance: Significance, *P* values and *t*-tests. *Nature Methods*, *10*, 1041–1042.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. doi:10.3389/fpsyg.2013.00863

Lakens, D. (in press). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*.

Lakens, D. (2014b). *Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses*. Retrieved from http://ssrn.com/abstract=2381936

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

Miller, J., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods*, *16*, 337–360.

Nelson, L. D., Simonsohn, U., & Simmons, J. P. (2014). *P-curve fixes publication bias: Obtaining unbiased effect size estimates from published studies alone*. Retrieved from http://ssrn.com/abstract=2377290

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301. doi:10.1037/1082-989X.5.2.241

Nosek, B. A., & Lakens, D. (in press). Registered reports: A method to increase the credibility of published results. *Social Psychology*.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

Proschan, M. A., Lan, K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York City, NY: Springer.

Reboussin, D. M., DeMets, D. L., Kim, K., & Lan, K. K. (2000). Computations for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials*, *21*, 190–207.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. doi:10.1037/1089-2680.7.4.331

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*, 293–304.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. doi:10.1016/j.jrp.2013.05.009

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, *55*, 62–71.

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*, 1359–1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L., & Simmons, J. (in press). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632.

Stanley, D., & Spence, J. (2014). Expectations for replications: Are yours realistic? *perspectives in Psychological Science*, *9*, 305–318.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.