AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2022 American Psychological Association
ISSN: 1082-989X

2024, Vol. 29, No. 5, 844–867
https://doi.org/10.1037/met0000520

Sample Size Planning for Replication Studies: The Devil Is in the Design

Samantha F. Anderson[1] and Ken Kelley[2]
[1] Department of Psychology, Arizona State University
[2] Department of IT, Analytics, and Operations, Mendoza College of Business, University of Notre Dame

Abstract

Replication is central to scientific progress. Because of widely reported replication failures, replication has received increased attention in psychology, sociology, education, management, and related fields in recent years. Replication studies have generally been assessed dichotomously, designated either a "success" or "failure" based entirely on the outcome of a null hypothesis significance test (i.e., $p < .05$ or $p > .05$, respectively). However, alternative definitions of success depend on researchers' goals for the replication. Previous work on alternative definitions for success has focused on the analysis phase of replication. However, the design of the replication is also important, as emphasized with the adage, "an ounce of prevention is better than a pound of cure." One critical component of design often ignored or oversimplified in replication studies is sample size planning, indeed, the details here are crucial. Sample size planning for replication studies should correspond to the method by which success will be evaluated. Researchers have received little guidance, some of which is misguided, on sample size planning for replication goals other than the aforementioned dichotomous null hypothesis significance testing approach. In this article, we describe four different replication goals. Then, we formalize sample size planning methods for each of the four goals. This article aims to provide clarity on the procedures for sample size planning for each goal, with examples and syntax provided to show how each procedure can be used in practice.

Translational Abstract

Replication is fundamental to science and has been receiving increased attention in psychology and related fields. Whether a replication study is deemed a "success" or a "failure" has typically been deduced from the result of a null hypothesis significance test, wherein results reaching conventional criteria for statistical significance are considered successful replications and nonsignificant results are deemed failures. Recently, scientists have been encouraged to consider other approaches, consistent with alternative goals for the replication study. However, study design, and specifically sample size planning, has often been absent from the discussion. Sample size planning should be consistent with the particular replication goal and analysis method that will be used. In the present article, we articulate four different goals for replication and we present formal sample size planning guidance for each goal. We include empirical examples and computer syntax to demonstrate each procedure in practice.

Keywords: replication, sample size, statistical power, accuracy in parameter estimation, meta-analysis

Supplemental materials: https://doi.org/10.1037/met0000520.supp

Psychology, sociology, education, management, and related fields are undergoing a "crisis of confidence" (e.g., Pashler & Wagenmakers, 2012, p. 528). The good news is that widely reported replication failures in the literature, most notably large-scale projects such as the Reproducibility Project Psychology (Open Science Collaboration, 2015), have focused attention on areas for improved methodological rigor and shed light on replication as a key component of the scientific process. Although replication has been seen as fundamental dating even back to Fisher (see Goodman, 2016), replication studies have historically been less publishable in premier outlets that value "novel" work (R. M. Lindsay & Ehrenberg, 1993; Makel et al., 2012). This is beginning to change, though, through several initiatives by major governing bodies and premier journals (D. S. Lindsay, 2015; Shrout & Rodgers, 2018). As replication research grows in perceived importance and frequency, it is critical that replication studies are designed and analyzed in a way that aligns with researchers' goals. Importantly, the goal of a successful replication study is not necessarily the same as the goal of a successful primary study.

This article was published Online First July 21, 2022.

Samantha F. Anderson https://orcid.org/0000-0003-4451-5295

Ken Kelley https://orcid.org/0000-0002-4756-8360

This research has not been previously disseminated. This article does not report empirical data analyses, and as such, was not preregistered. All study materials have been made available. All demonstrated software is freely available, and access links are provided in footnotes. The R code to reproduce Figure 1 is included in the online supplemental materials.

Correspondence concerning this article should be addressed to Samantha F. Anderson, Department of Psychology, Arizona State University, 950 South McAllister Avenue, Tempe, AZ 85281, United States. Email: samantha.f.anderson@asu.edu

Recent work has posited thoughtful arguments for how to define success and failure in replications, with some suggesting multiple definitions are necessary depending on the research goals (Anderson & Maxwell, 2016; Bonett, 2021; Schauer & Hedges, 2021; Verhagen & Wagenmakers, 2014). Replication studies have generally been assessed dichotomously, designated either a "success" or "failure" based entirely on the outcome of a null hypothesis significance test (NHST; i.e., $p < .05$ or $p > .05$, respectively; see Hedges & Olkin, 1980; Maxwell et al., 2015). In addition to the historical and recent criticisms of NHST (e.g., Bakan, 1966; McShane, Gal, et al., 2019; Rozeboom, 1960; Wasserstein et al., 2019), defining success based on the outcome of the null hypothesis (H_0) can be overly narrow or ill-aligned with the goals of the replication. For example, Maxwell et al. (2015) describe a hypothetical replication, which aims to support the null hypothesis but relies on standard NHST, yielding results that convey an equal likelihood of the true effect size being zero or medium in magnitude. Thus, it is useful to consider other approaches, which are often preferable and will align more with researchers' replication goals, including definitions focused on null effects, estimation, and combining multiple studies. The extant literature on these alternative definitions has generally focused on the analysis of the replication study (e.g., Anderson & Maxwell, 2016; Schauer & Hedges, 2021; Verhagen & Wagenmakers, 2014). However, given the link between replication "failures" and statistical power and accuracy in parameter estimation (see Anderson & Maxwell, 2017; Gilbert et al., 2016), it is vital to consider how to appropriately *design* replication studies, in which consideration of sample size planning is critical. Indeed, the devil is in the details.

Sample size planning, even for an original study that intends to use a standard analysis technique, can be challenging. The process can become even more confusing for researchers conducting replications, particularly if they want to approach replication in sophisticated ways. The extant literature offers little guidance, as sample size planning most appropriate for several replication goals either does not exist or is (a) scattered throughout the methodological literature, (b) not clearly linked to replication studies, or (c) focuses on technical or computational elements, rather than bigger-picture issues involved in the process of sample size planning. In particular, there is no single source that discusses sample size planning, let alone provides methods for implementation, for replication studies outside of the most limited approach focusing on NHST. It is striking that so little attention has been paid to linking effective sample size planning for different replication goals in practice, such that researchers can understand the process involved in planning a well-designed replication study under a particular goal. This article fills this void and provides thorough—and practical—guidance on sample size planning for four important replication goals.

We have three primary aims with this article. First, we aim to augment the literature by developing a coherent literature review that builds a bridge between sample size planning literature and the replication literature, integrating diverse literatures that most researchers would not usually have in their corpus of readings, and explicitly discussing how to plan replication studies under a variety of goals suggested for replication. Readers may be familiar with analyses for some of these replication goals, but we detail important insights relevant to the planning phase, which has received considerably less attention. Similarly, readers may be aware of some of these sample size planning approaches, although many are underused even outside of the realm of replication and have not been detailed in an accessible way nor linked with replication.

Second, we aim to diverge from the methodological sample size literature somewhat, which has often been more technical or computational, to focus on broader questions that arise when planning for replication studies in practice, such as how to conceptualize the effect size, how much to rely on the original study, and whether the burden of proof should rest on the null or alternative hypothesis. In line with this, we avoid highly technical mathematical and programming details, which can be found in the methodological literature on sample size planning cited herein. In fact, we contend that the coding necessary to use most sample size planning packages, including those we demonstrate here, is quite simple, relative to broader issues in deriving the inputs, and we provide much of it. We believe that this presentation will expand the userbase and thus bring useful methods to a more diverse audience.

Third, we aim to demonstrate sample size planning in practice for each goal. We use empirical examples and provide computer code using freely available software so that the methods discussed can be immediately implemented by researchers. Chiefly, we focus on standard statistical models, including analysis of variance, *t*-tests, and regression procedures. These standard analyses are (a) exceedingly well-represented in replication studies, given evidence from large-scale replication projects and the subfields for which replication is most represented,[1] and (b) permit us to focus more on process-oriented details and concepts. However, we provide citations to more technical sources for interested readers.

Definitions and Replication Goals

One of the difficulties of interpreting replication studies is that definitional terms in the replication literature are often inconsistently defined, used interchangeably, or are not defined at all (Bonett, 2021; Hedges & Schauer, 2019b). Replication researchers sometimes mix multiple different analysis strategies, thus conflating the interpretation of the replication study and leading to unclear methods for planning such studies (Anderson & Maxwell, 2016; Schauer & Hedges, 2021). A variety of taxonomies have been developed to differentiate types of replication based on the goals of the replication study. First, reproducibility needs to be differentiated from replicability. Reproducibility is "the ability to recompute data analytic results conditional on an observed data set and knowledge of the statistical pipeline used to calculate them," whereas replicability is "the chance that a new experiment [or study more generally] targeting the same scientific question will produce a consistent result" (Patil et al., 2016, p. 540). The focus here is on the replication, as reproducibility does not require new data and represents an entirely different aspect of science. Moreover, we use the term replication broadly to refer to the process of conducting a replication study, which may or may not yield results that successfully replicate by a particular definition.

Second, replications can either be direct or conceptual, with the former aiming to mimic the original study conditions closely and

[1] For example, based on our examination of RPP studies, only 2% used a focal analysis other than ANOVA, *t*-tests, linear regression/correlation, and binomial or χ^2 tests.

the latter aiming to expand the original study using new methods or to generalize results to a new setting or condition (Schmidt, 2009; Simons, 2014). Although there is debate as to whether direct (e.g., Simons, 2014) or conceptual (e.g., Crandall & Sherman, 2016) replications are more important, it is clear that a research literature involving both formats can be beneficial, depending on what goals are most relevant to a particular field at a particular point in time (including replications that blur the line between exact and conceptual; "systematic replication," Hendrick, 1990, p. 41).

The focus of this article will be on direct replications, which have been a primary interest in the literature and represent an ideal context to develop "useful, testable, and generalizable predictions" for a phenomenon (Simons, 2014, p. 79). That said, we note that even direct replications should typically not be expected to *exactly* replicate the original study *results* (see McShane, Tackett, et al., 2019). This fact has direct relevance for a commonly desired, logically sensible, but ultimately too stringent replication definition that centers on the results rather than the process used: that the replication effect size exactly replicates the original, in the sense that the effect sizes are identical (exact replication; Hedges & Schauer, 2019b). First, even a perfectly identical replication study in terms of procedures and measures is subject to sampling error, namely the estimation variability from study to study, even if nothing has changed and the population is the same. Second, when invoking the degree of similarity between the original and replication study conditions, the issue of hidden effect variation can arise, which is variability due to seemingly minor contextual differences, and effect size drift, which is that the nature of the effect may change over time. McShane and Bockenholt (2014) go so far as to argue that "you cannot step into the same river twice" (p. 612), where the idea is that effect sizes are not fixed (and therefore that we should not expect exactly the same results). Indeed, if research is done to unlock the current state of affairs of nature, and the state of affairs or nature changes, how could one expect the effect size to be exactly the same value? These multiple sources of heterogeneity, even for the most carefully monitored replication studies, support more approximate replication, which allows for some level of variation (Hedges & Schauer, 2019b). Thus, a "perfect" or "identical" replication result should not necessarily be expected, even in direct replications (Bonett, 2021; Kenny & Judd, 2019; McShane & Bockenholt, 2014; Stanley & Spence, 2014). In the end, there is a philosophical debate as to what degree of exact replication is achievable or desired, or whether the effect size of interest is changing, either due to characteristics of the population or the particulars of the conditions (even if the changes are unknown or thought to be immaterial to the research). We argue that setting a customized goal and appropriately planning sample size accordingly can, at a minimum, reduce excess variability when a more exact replication is desired and more accurately identify what conditions influence the nature of an effect when a more approximate replication result is expected.

Third, replication researchers may have different underlying goals for what they would consider constitutes a *success* or *failure*, which should imply selecting different analyses to achieve those goals. Relatedly, it is important to distinguish between the method used to analyze the data in the replication study (e.g., a between-subjects factorial ANOVA) and the method used to ultimately decide if the replication study replicated or failed to replicate the original findings (e.g., a confidence interval [CI] for the average effect size). These distinct constructs sometimes overlap (e.g., such as when a p-value from an ANOVA is used to decide a replication's success, as described in the next paragraph), but not necessarily.

Anderson and Maxwell (2016) found that most researchers conducting replication studies analyzed their data in the same way as the original study. In particular, for an original study that had achieved statistical significance, researchers commonly decided that a replication was successful on the basis of a statistically significant p-value from the same analysis method and that a replication was a failure on the basis of a nonsignificant p-value using the same analysis method. This finding raises two focal points. First, using the same analysis method as the original study is appropriate in some cases, but using a (possibly impoverished) method simply because the original researchers did should not be the exclusive means of conducting replication. In fact, this degree of conformity can even be detrimental. For example, take the case of a randomized treatment study. If a covariate was collected at baseline for a randomized design, ANCOVA would provide a more powerful test in the replication study of the same question of evaluating mean differences across groups if the covariate is sufficiently correlated with the outcome variable, even if the original study used ANOVA (see Chapter 9, Maxwell et al., 2018). And for a longitudinal study, a mixed-effects model might be preferred under certain conditions to a repeated-measures ANOVA, even if the original study used the latter.[2] Second, regardless of the type of analysis used, reliance on NHST and p-values to decide replication success should not be the default approach. In addition to issues we raise later, in Anderson and Maxwell's (2016) review, replication researchers sometimes focused on p-values when their stated goal would not imply such an approach, and in other cases, did not seem to be considering that replication success could be defined more broadly than within the confines of NHST.

Methodologists have thus expanded replication research to feature several goals. We present a comprehensive list of these recently proposed goals in Table 1, which arises out of an emerging literature that continues to evolve as researchers think critically about replication. Some of the goals put forth focus on establishing the existence or direction of an effect, others focus on investigating differences in effect magnitude, and still others use the advantage of multiple replication studies to conduct meta-analytic tests for pooled effect sizes and heterogeneity. Importantly, the extant literature has focused mainly on how to define successful replication and analyze replication studies, rather than on the design of the replication. Appropriate sample size planning should take the goal and intended analysis strategy into account, otherwise there is a mismatch between design and analysis.

[2] A related, but tricky point, is what to call ostensibly direct replications that make improvements to the analysis approach. Some of these replications may be testing slightly different statistical hypotheses (e.g. as tested by ANOVA versus ANCOVA) although this is not always the case (e.g., using a separate-variance approach to adequately handle violated assumptions). Among the categories of direct and conceptual, we would conjecture that these are still in the spirit of direct replication, but ultimately considering direct and conceptual replication as more of a continuum may prove more effective in the long term.

Table 1
Replication Goals From Various Taxonomies

Goal	Reference
Evidence for existence/absence of effect	
To infer the existence of a replication effect	Anderson & Maxwell, 2016
Directional replication	Bonett, 2021
To infer a “null” replication effect	Anderson & Maxwell, 2016
“Null” effect size nonreplication	Bonett, 2021
Is the effect present/absent in the replication attempt?	Verhagen & Wagenmakers, 2014
When pooling all data, is the effect present or absent?	Verhagen & Wagenmakers, 2014
Inconclusive directional nonreplication	Bonett, 2021
Magnitude estimation/difference of effect sizes	
To precisely estimate the replication effect size	Anderson & Maxwell, 2016
To assess whether replication is clearly inconsistent with original	Anderson & Maxwell, 2016
Weak effect size nonreplication	Bonett, 2021
To assess whether replication is clearly consistent with original	Anderson & Maxwell, 2016
Does the effect size in the replication attempt equal the effect size in the original study?	Verhagen & Wagenmakers, 2014
Strong effect size nonreplication	Bonett, 2021
Strong effect size replication	Bonett, 2021
Inconclusive strong effect size nonreplication	Bonett, 2021
Inconclusive strong effect size nonreplication	Bonett, 2021
Does the original study effect size come from same distribution as replication study effect size?	Mathur & VanderWeele, 2019
Do the original and replication study confidence intervals overlap?	Brandt et al., 2014
Pooled/meta-analytic approaches	
To combine replication sample data with original results	Anderson & Maxwell, 2016
Continuously-cumulating meta-analysis for replication	Braver et al., 2014
Meta-analytic test for exact (effect parameters match) replication with burden of proof on nonreplication	Hedges & Schauer, 2019b
Meta-analytic test for approximate (practical equivalence) replication with burden of proof on nonreplication	Hedges & Schauer, 2019b
Meta-analytic test for approximate (practical equivalence) replication with burden of proof on replication	Hedges & Schauer, 2019b
What effect would we expect to see in the replication study once we have seen the original effect (prediction interval)?*	Patil et al., 2016
Proportion of replication effects larger than minimum scientifically meaningful effect size	Mathur & VanderWeele, 2019
Combine meta-analysis with preregistered replication	Carter & McCullough, 2018
Other goals	
Do the replication data support a skeptic’s (null) or a proponent’s (idealized belief based on original study) hypothesis more?	Verhagen & Wagenmakers, 2014
Directional nonreplication	Bonett, 2021

Note. Goals in italics are equivalent or essentially equivalent to the approach directly above.
* Equivalent to meta-analytic Q test with $k = 2$ studies.

This article delineates four general sample size planning approaches for replication studies and provides solutions for their implementation:

1. To infer the existence of a replication effect;
2. To infer a “null” replication effect;[3]
3. To precisely estimate the magnitude of the replication effect size;
4. To combine results across one or more replication studies.

This list of goals is certainly not exhaustive (see again Table 1); however, we elect to explicate these four because (a) they represent what our experience suggests are meaningful, and varied, approaches for replication; (b) motivation for these goals as relevant to replication has been established in prior literature (see e.g., Anderson & Maxwell, 2016; Bonett, 2021; Maxwell et al., 2015); (c) there are no technical flaws in the underlying logic for these goals (unlike, e.g., CI overlap; see Schenker & Gentleman, 2001); (d) they map onto broader discussions within the social sciences on how to generate cohesive literatures using hypothesis testing, equivalence testing, magnitude estimation, and meta-analysis. Readers may, however, notice that a seemingly attractive goal is missing, one which answers the question of whether the original and replication study effect sizes are equal, which can be done using a hypothesis test (meta-analytic Q-test) or a CI for a difference in effect sizes. Indeed, analysis procedures to achieve this goal have been described (Anderson & Maxwell, 2016; Bonett, 2009; Schauer & Hedges, 2021). However, given that (a) this uses an often unattainable exact definition of replication due to multiple sources of heterogeneity (see our previous discussion & Mathur & VanderWeele, 2019); and (b), statistical power and accuracy of the estimated parameters for this is highly limited by the sample size of the original study (see Bonett, 2021), we consider sample size planning for alternative goals here (though, see Schauer &

[3] We use quotations surrounding “null” to emphasize that this goal will focus on showing an effect is so trivial as to be considered null, rather than to prove the effect is literally zero. In some cases, this subgoal of replication could be termed “refutation,” but as we detail later, the goal is broader than simply refutation.

Hedges, 2021; for an extension that considers heterogeneity among effect sizes from multiple studies).

For each of the four goals above, we provide sample size planning guidance and methods within the classical paradigm of frequentist statistics, given that the majority of studies, and replications specifically, rely on such classical methods. However, we emphasize that the goals are agnostic to the overarching paradigm, and we note Bayesian procedures in the spirit of each goal. We also introduce Bayesian sample size planning in the context for replication toward the end of the article. Although these sections are certainly not detailed enough for readers to conduct fully Bayesian sample size planning or integrate such procedures into Bayesian replications, our overarching goal is to help researchers think more flexibly about replication, and we expect that additional goals, procedures, and guidance will be proposed to achieve this flexibility in the Bayesian paradigm. In fact, an analog of this article focused on Bayesian methods would be useful. We hope that this article encourages these goals and variations of these goals to be more fully considered by researchers because it is clear a more nuanced approach to replication is needed. Importantly, we encourage readers to appreciate the variety of thoughtful proposals shown in Table 1 (e.g., Bonett, 2021; Brandt et al., 2014; Braver et al., 2014; Carter & McCullough, 2018; Mathur & VanderWeele, 2019; Schauer & Hedges, 2021), and we consider other approaches in the discussion section that may be useful.

Statistical Power and Accuracy in Replication

Sample size planning is designed to achieve calculable parameter estimates (e.g., the quantity can be calculated, the model will converge) at a minimum, paired with a desired level of statistical power or accuracy. Statistical power is the probability of detecting an effect at a fixed (unknown) size, using a significance test at the specified α level, assuming a specified point value for the null and a nonzero (fixed) effect. To clarify, for a composite alternative hypothesis, power at a particular sample size and α level has a distribution, but reduces to a single value when conditioning on a fixed population effect size (Lehmann & Romano, 2005). Accuracy is a function of precision and bias. Precision is the reciprocal of variance, and the more precise an estimate the narrower its corresponding CI for the population effect, holding everything else constant, such that the sample estimate is contained within a narrower set of plausible parameter values. Bias is a systematic difference between the expected value of an estimate and the value of the relevant population parameter. Accuracy and precision are equivalent when the estimate is unbiased (Kelley & Rausch, 2006, p. 363). Given that researchers ultimately desire accuracy when possible, we generally use "accuracy" throughout, except when precision specifically is of note.

Methodologists have articulated the distinctiveness of power and accuracy not as "rivals" (e.g., Kelley & Maxwell, 2003; Maxwell et al., 2008, p. 308). Rather, the aim of power or accuracy must depend on the theoretical framework that the study is situated within and, importantly, the goals of the study. In some domains, typically more applied settings (e.g., health, clinical psychology, organizational behavior/management), there are clear clinical, practical, or managerial implications of certain effect sizes (e.g., Beaton et al., 2002; Paterson et al., 2016). In these circumstances, estimating the effect size may be the focal objective and therefore sample size planning for accuracy is most appropriate, so as to accurately infer the magnitude of the phenomenon. Other studies are designed to test a theory, compare competing theories, or find the direction of the effect (e.g., Festinger & Carlsmith, 1959; Nordgren & McDonnell, 2011). In some of this work, the existence and/or direction of the effect is focal, rather than its magnitude (Dawes, 2004; Haaf et al., 2019; Morey et al., 2014), and sample size planning for statistical power may be more appropriate. Taking a causal inference perspective, West and Thoemmes (2010) noted a parallel dichotomy between estimating direction and magnitude and clarified that the "strengths of the two perspectives are often complementary" (p. 19; consistent with the idea they are not rivals, as we noted above). We share this perspective that various methods of sample size planning can coexist, and we agree with Kruschke and Liddell (2018), who noted "the goal determines how we plan the research" (p. 202). Some of the replication goals described in this article lend themselves more naturally to significance testing approaches and thus sample size planning for power will be considered. Other goals lend themselves more naturally to accuracy-based approaches.

Though replication studies are generally a new phenomenon in the literature, they have tended to be underpowered to address their focal goals, even when using larger samples than the original study (Shrout & Rodgers, 2018; see also Anderson et al., 2017). This lack of statistical power for the test of the replication (i.e., replication power; Held et al., 2020) is not always caused by a lack of attention to sample size planning, rather, sometimes well-intentioned researchers simply use an approach that does not acknowledge, for example, publication bias and/or uncertainty and their implications on the replication. In particular, many approaches rely on imperfect information from the original study, as we will demonstrate graphically in our subsequent discussion of Goal 1. Because of this, replication power using various approaches often depends on that of the original study. Related to this, some methodologists (Hedges & Schauer, 2019a; Maxwell et al., 2015) have issued a warning about comparing one or multiple replication studies to a single original study, which we summarize as *unless the original study has a high degree of statistical power, with accurate parameter estimates, the statistical power of the test of replication is often constrained by the power of the original study and the accuracy of the parameter estimates therein*. In a study on NHST-based replication, researchers found that traditional methods of sample size planning, which might claim to provide 80% power in the replication study, could provide abysmally low power when the original study was itself underpowered, even if original study yielded a statistically significant finding (Anderson & Maxwell, 2017). This phenomenon occurs when sample size planning relies on the original study's sample effect size reported in the article, which is often *larger* than the *true* value, leading to a replication sample size that is unintentionally *smaller* than it should be. In other words, the sample size is calculated assuming a larger effect magnitude than exists in reality, resulting replication studies underpowered for key statistical tests. This is a key point that has been misunderstood. Based on the logic presented here and probabilistic outcomes, we argue that if original and replication studies were appropriately powered, there would be many more studies—though certainly not all—that would replicate original studies' findings, at least by one definition (i.e., achieve statistical significance, in the same direction).

Because of the issues we have raised about statistical power in the original study constraining statistical power in the replication study if the issues discussed here are ignored, we discuss sample size planning from the meta-analytic perspective, which has a long history in research synthesis more generally, but has recently been advocated for replication (e.g., Anderson & Maxwell, 2016; Braver et al., 2014; Fabrigar & Wegener, 2016; Hedges & Schauer, 2019a). A meta-analytic perspective is advantageous when multiple replication studies on a topic are available (e.g., planning the i^{th} replication based on an original study and $i - 1$ prior replications). An important question, then, is whether there is value in describing sample size planning for single replication attempts at all. We argue that there are reasons to include guidance for single replication studies, which we now discuss.

First, despite the attention given to multisite replication projects (e.g., Many Labs, Klein et al., 2018; CREP, Wagge et al., 2019), the majority of replications tend to be single, piecemeal attempts at replication. Even large-scale replication projects often include a single replication attempt per original study (e.g., Replication Project Psychology, Open Science Collaboration, 2015) to prioritize the number of studies replicated rather than the number of replications or compare multiple replication studies to a single original target study, which does not necessarily ameliorate the aforementioned issue (Schauer & Hedges, 2020). Further, even when the eventual research goal is to conduct multiple replications, only a single replication attempt is usually available to plan at the outset of this process. Such a replication should have an appropriate sample size so that a sufficiently powerful test or precise estimate for the initial conclusion regarding replication success can be made. Second, despite the fact that the literature is filled with studies underpowered for their focal research questions (Bakker et al., 2016), and therefore original studies that are selected for replication themselves have underpowered tests, there are also high-powered tests in original studies, especially with certain types of samples (e.g., registries; Gliklich et al., 2014), methods of data collection (e.g., MTurk; Necka et al., 2016), and subfields (e.g., health psychology; Maddock & Rossi, 2001). An optimistic researcher might believe that the recent increase in attention to sample size planning, which was at least somewhat driven by the replication crisis, may markedly improve the power and precision of original studies. In some ways original and replication studies share common problems, and improvements in design will positively influence both types of studies and lead to a more cohesive literature. We do caution, though, that a somewhat skeptical approach to results from single replications may be warranted as researchers await more replication studies and meta-analytic summaries.

Sample Size Planning Strategies for Four Replication Goals

We now summarize four replication goals and provide solutions for researchers with one of the goals. The collection of these four goals are what we believe to be the most commonly sought-after replication approaches. We then provide a solution so that not only do researchers understand what they might want to do, but so they can actually plan the needed sample size under a more realistic scenario. In particular, we make available extensive resources for researchers that combine the theoretical background and the numeric implementation into easy to use R packages, and web apps. We believe that our discussion as well as computational resources will help advance not only the framing of the arguments but also the application to practice. We provide a guiding framework for replication advancement, and, importantly, also the necessary tools for implementation.

Goal 1: To Infer the Existence of a Replication Effect

The first goal is when interest concerns inferring the existence of an effect via a replication study, where existence is operationalized based on the results of a hypothesis test. In the classical paradigm, replication studies are commonly analyzed via NHST, using standard statistical techniques matched to the previous analysis, an analysis approach which may be sensible when Goal 1 is of interest. A replication is often declared "successful" when it achieves a statistically significant p-value, assuming the original study achieved significance and typically assuming the direction or sign of the effect matches the original finding (Anderson & Maxwell, 2016; Bonett, 2021).

Given limitations of NHST and issues that can be compounded by NHST, such as the garden of forking paths (Gelman & Loken, 2014), HARKing (Kerr, 1998), and questionable research practices (Anderson, 2020; Manapat et al., 2022; Simmons et al., 2011), we argue that this metric for determining replication success is used more often than it should be. On the one hand, a successful replication using standard NHST may have an effect that is much different in magnitude than reported in the original study, limiting its usefulness when estimating magnitude is fundamental. On the other hand, a failed replication using NHST is not evidence that the original finding was a false positive. Researchers have described confounding factors such as low statistical power in the original and/or replication study (e.g., Anderson & Maxwell, 2017), heterogeneity among effects (e.g., Kenny & Judd, 2019), and sampling error (e.g., Stanley & Spence, 2014). Furthermore, as Gelman and Stern (2006) demonstrated, the difference between statistically significant and nonsignificant is itself not necessarily statistically significant. Given, these issues and some we develop later, we encourage readers to consider some of the alternative goals and associated strategies we discuss, or to adopt a Bayesian interpretation of Goal 1, relying on properties of the posterior distribution or various versions of the Bayes Factor (Haaf et al., 2019; Schönbrodt & Wagenmakers, 2018; Verhagen & Wagenmakers, 2014).[4]

However, we still believe it is worthwhile to discuss sample size planning within the context of NHST for a few reasons. First, despite its criticisms, current methodological viewpoints regarding NHST and associated p-values are varied and nuanced (see e.g., American Statistical Association Task Force, 2021; Greenland, 2019; Kafadar, 2021; Kuffner & Walker, 2019; Mayo, 2020,

[4] There are multiple ways in which replication has been operationalized within the Bayesian approach that are in keeping with the focus of Goal 1. Some have suggested novel forms of the Bayes Factor (e.g., Schönbrodt & Wagenmakers, 2018; Verhagen & Wagenmakers, 2014), although there is currently some disagreement among Bayesian methodologists regarding the Bayes Factor and Bayesian hypothesis testing in general (Tendeiro & Kiers, 2019; van Ravenzwaaij & Wagenmakers, 2021). A reviewer suggested that the posterior could be used in the spirit of Goal 1, answering questions such as "what is the probability that the effect is positive?" or "what is the probability that the effect is positive and above a certain minimum effect size?"

2021). Second, as we have explained, replication with the goal of showing the existence of an effect, even relying on NHST, may have merit in studies strongly linked to theory building, where existence and direction of effects are often more focal than size. Correspondingly, a NHST-based procedure may be appropriate and align with Goal 1, although of course this does not imply that other approaches could not also be motivated here. Third, even if all methodologists came to agree that NHST is always inferior, replication studies using NHST will continue to populate the literature for some time, and we believe there is still much to be gained from making improvements within this paradigm. Finally, some of the issues we detail within the section that follows, such as publication bias, are relevant across other perspectives.

In the classical framework, sample size planning for Goal 1 is perhaps the most familiar of the goals described here, because it is treated similarly to, but importantly it is not identical to, a typical a priori power analysis for a nonreplication study. The aim of sample size planning for this goal is to achieve a desired level of statistical power for the replication study. Until recently, sample size planning for replication studies using standard NHST techniques under Goal 1 typically yielded lower than desired power, because it failed to consider that the original study's effect size contains publication bias and sampling error, in addition to the between-study heterogeneity described in the Introduction. Given that an effect size is a necessary input in standard sample size planning software (as an indirect measure of the noncentrality parameter) and given that the original study may provide effect size information of (seeming) direct relevance to the hypothesized replication effect, researchers have often used the sample effect size reported in the original study as the value for the effect size input in a priori replication sample size planning.

Yet, standard methods of sample size planning assume that the hypothesized effect size is a population value, which is almost certainly not the case. Furthermore, as we have discussed, the average of a set of published values will themselves not generally be the population value. Sample effect sizes reported in published studies are subject to publication bias (Lane & Dunlap, 1978) and uncertainty (Dallow & Fina, 2011; see also, McShane & Bockenholt, 2016; O'Hagan et al., 2005; Perugini et al., 2014), and will therefore generally—if not always—differ from their corresponding population values. Publication bias results from the fact that studies reporting effects larger in absolute value, which may differ from zero due to sampling error, will have a smaller *p*-values and thus are more likely to be statistically significant and consequently published, due to journal preference for statistically significant results and investigator belief in that preference (Brown et al., 2017), than studies reporting smaller effects. In other words, the published literature does not reveal the entire distribution of effects, often what is missing are those that do not meet the specified criteria for statistical significance (i.e., $p < \alpha$). Practically speaking, this results in published study effect sizes being *larger* in magnitude than the true effect size (i.e., there is a positive publication bias—leading to an inflated perception of the true value in the population). Uncertainty adds an additional layer of complication, in that the sample effect size is an imperfect estimate of the population effect size due to sampling error, which is best illustrated via a CI for the population effect size, where a wide CI illustrates the uncertainty with which a population effect size has been estimated. Thus, when unadjusted sample values are used in planning replication study sample sizes for Goal 1, the resulting sample size will tend to be too large and thus the actual statistical power will be lower than the nominal statistical power that is often reported in the replication by the researcher. To make matters worse, the association between effect size and power is not linear: the power loss from using an overestimate is worse than the power gain from using an underestimate (Maxwell et al., 2015), which is based on the same principle as power curves.

To illustrate publication bias in practice, consider a situation in which the population value of the standardized mean difference, δ, for a test of two independent means is $\delta = 0.2$. Suppose that 100 studies are conducted, each with a fixed sample size of $n = 30$ per group. With this sample size, suppose that 12 studies find statistically significant effects.[5] Thus, under this model of publication these 12 studies will be "publishable." Although the average sample value of Cohen's *d* among all 100 studies is very near .20, (i.e., close to the population value), the average value of Cohen's *d* from the published studies is .61, which represents a bias of .41 (= .61 − .20). This is illustrated in Figure 1 (reproducible code in online supplemental materials), where the mean of the 12 published studies (solid line) is much higher than the overall mean (dotted line) of both published (black triangles) and unpublished (gray circles) studies.

Consequently, a future researcher may use one of these overly large effect size estimates—we will demonstrate with the average of the 12 publishable studies—and plan sample size for the replication based on such a power analysis. Doing this, which has been recommended—not considering publication bias—leads to an assumed effect size value used in the sample size planning procedure that, correspondingly, overestimates the magnitude for the replication, thereby leading to a smaller than needed sample size. That is, the researcher ends up with an underpowered study to detect the effect. This replication study would be unlikely to reach statistical significance, and many would consider such an outcome a "failed" replication. In particular, rather than power being about 11.54% as it actually is for a sample size of 30 in this scenario, a power analysis based on the published studies would be estimated to be 64%.[6] Further, using $d = 0.61$ as if it were the true value, a power analysis would suggest the sample size per group needs to be $n = 44$ due solely to the highly inflated estimate of the true effect from the literature.[7] Using a sample size of $n = 44$ per

[5] The power in this situation is about 12%. In other words, we would expect about 12 studies to be statistically significant. This is evident via running the following R code:

```
power.t.test (delta = .2, sig.level = .05,
  n = 30).
```

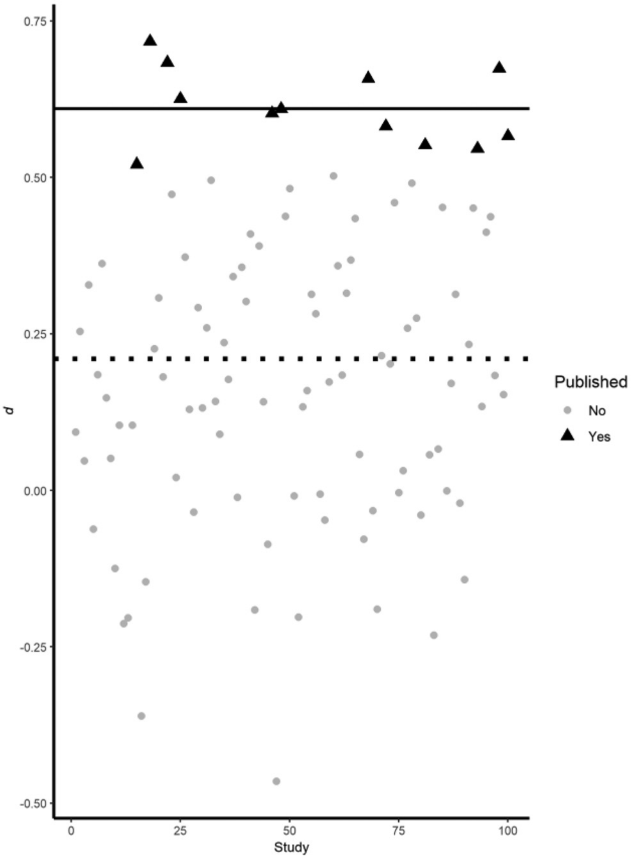
[6] This can be demonstrated with the following R code:

```
power.t.test (delta = .61, sig.level = .05,
  n = 30).
```

[7] The sample size seemingly needed can be seen with the following R code:

```
power.t.test (delta = .61, sig.level = .05,
  power = .80).
```

Figure 1
Example Illustrating the Influence of Publication Bias on Effect Sizes



Note. The figure uses 100 hypothetical original studies, where the population standardized mean difference is δ = 0.2 and the per-group sample size of all studies is 30. The *x*-axis reflects the study number. The *y*-axis reflects the sample effect size, *d*. The 12 sample effect sizes calculated from studies attaining statistical significance are shown as black triangles ("published"). The other 88 sample effect sizes are shown as gray circles. The lower of the two horizontal lines (dotted) indexes the mean *d* of all 100 studies, which is 0.21. The upper of the two horizontal lines (solid) indexes the mean *d* of only the 12 published studies, which is 0.61.

group in a situation where the true $\delta = 0.2$ leads to just over 15% power for the replication study.[8] This is why we believe that many "failed" replications involve a failure of design: The effect may well be nonzero but less detectable in replications due to sample size planning that does not account for publication bias.

There are two general ways to ameliorate these issues. First, researchers could simply ignore information from the original study and use a theoretical population effect size value as input in sample size planning. This approach is described at the end of the current section. Alternatively, researchers can use information from the original study that has been adjusted for publication bias and uncertainty. Perugini et al. (2014) developed a method that adjusts for uncertainty by using the lower limit of a two-sided CI of a percentage chosen by the researcher, with larger percentages correcting for more uncertainty. The researchers recommend a 60% or 90% CI, though these choices are more or less arbitrary as to what confidence level should be used. Hedges (1984) derived the distribution for δ under conditions of publication bias, and the formula can be used to adjust two-group studies for publication bias, but not uncertainty. Another approach is the bias-uncertainty corrected sample size (BUCSS), which is a framework that adjusts sample effect sizes for both uncertainty and publication bias.[9] The underlying mathematical basis and software (Anderson & Kelley, 2020) are described elsewhere (e.g., Anderson et al., 2017; Anderson & Maxwell, 2017; Taylor & Muller, 1996). However, a brief description and software overview are provided here, with an example, so that this article can serve as a broadly applicable reference for sample size planning for replication studies.

To adjust for publication bias, the BUCSS framework relies on a truncated distribution to find the most likely true value of the original study effect size, where the truncation occurs at the suspected level of publication bias. The median of the truncated distribution of possible effect sizes is used as the *most likely* value when adjusting only for publication bias. Researchers can adjust for uncertainty by selecting a more conservative value from this distribution. This selection is made through the researcher's choice of an assurance value, which quantifies the desire to obtain the desired results realizing the fallibility of the input parameters. More specifically, assurance is used to represent the proportion of times the sample size planning method will result in a replication study with the desired level of power or higher. This involves considering the long run likelihood that a particular approach will meet or exceed its desired power for the replication study. For example, assuming the researcher plans for 80% power, selecting 50% assurance implies that 50 out of every 100 (hypothetical) studies will achieve 80% power or better. With respect to the truncated distribution that forms the basis for BUCSS, choosing higher assurance (> 50%) tells the program to select a lower, more conservative, value of effect size than the median. Specifically, the quantile is selected at the (1 – *assurance*) percentile. Assurance is described in much greater detail in other work (Anderson, 2021; Anderson et al., 2017).

We developed the BUCSS R package to implement this goal (Anderson & Kelley, 2020), which is also available with accompanying web apps (see DesigningExperiments.com) to more easily implement the method discussed here without any coding or knowledge of R (users would not necessarily know that R is used on the back-end of the web apps). The BUCSS R package and web apps accommodate many designs in the general linear model context (e.g., various types of ANOVA and linear regression). For ANOVA, the software supports main effects, interactions, and contrasts in between-subjects, within-subjects, and mixed ANOVA with any number of factors and levels. For regression,

[8] The true power using the population effect size and the suggested *n* can be seen with the following R code:

```
power.t.test(delta = .2, sig.level = .05,
  n = 44).
```

[9] BUCSS considers sampling error, but does not consider other forms of uncertainty or effect size variation. See McShane and Bockenholt (2016) for sample size formulas that directly consider between-study variation in effect sizes. See also Pek and Park (2019) for a Bayesian-Classical hybrid sample size planning procedure that considers multiple types of uncertainty.

the software supports simple regression and multiple regression, tests of a single predictor, as well as tests of multiple predictors jointly. Essentially, the BUCSS R package and web apps implement the traditional and still most widely used methods in many areas, though certainly not all areas, of psychology and related disciplines. Rather than requiring a specific type of effect size input, the software requests the observed *t* or *F* statistic, more universally reported in published studies, and then appropriately links this value with the noncentrality parameter, a quantity that combines effect size and sample size and directly relates to the degree to which the null hypothesis is false (and therefore statistical power). In addition, researchers will need to enter the original study sample size, their assumed level of publication bias, and their desired levels of statistical power and assurance. BUCSS outputs the suggested planned study sample size and the adjusted previous study noncentrality parameter based on the likelihood methods discussed here, which can be transformed to various effect size metrics and used as an input in other types of sample size planning if a publication bias and/or uncertainty-adjusted estimate is needed.[10]

Before illustrating the use of BUCSS, a few notes are needed. First, BUCSS assumes a model of publication bias wherein original study results are only available to be used for replication when $p < \alpha$. In this case, α refers to the maximum *p*-value needed for the result to be available to the researcher to replicate. This is a simplification of reality, wherein publication bias may be more complicated, but the assumption is valid in many research contexts, due to the literature often not considering studies that are do not have statistically significant findings for the principle outcomes, and the researcher can select what α is most sensible to their specific situation (e.g., .05 for a focal original result published in a standard journal). Larger α values are appropriate for situations where publication bias is thought to be less extreme and assume a lower degree of effect size inflation (e.g., replicating a registered report). Second, for *t*-tests, where either two-sided or one-sided tests can be performed, BUCSS assumes a two-sided test for the replication study. Although some may prefer a one-sided test for replication, given that the effect has a proposed (original study) direction, we prefer a two-sided test so that the replication study is better equipped to show an effect in the opposing direction, as would be needed, for example, in the case of a type-S error (Gelman & Carlin, 2014). When using the *t* test functions, replication researchers who wish to plan for a one-sided test can simply double the planned type I error level they enter into the software (e.g., if the plan is to conduct a one-sided test at $\alpha = .05$, one can enter an of $\alpha = .10$).

Third, an issue may arise when the sample size planning approach cannot be accomplished with the input parameters because the combination of factors is asking too much from the data: There simply is not enough information available. In particular, the researcher is asking for more certainty than the original study's information provided allows, at the researcher's desired levels of power, assurance, and publication bias adjustment. Researchers can adapt these values to make sample size planning feasible, particularly if the initial intended adjustments were more extreme (i.e., 95% assurance). Conceptually, this reflects one of the cases where an original study's power can limit that of the replication study: the original study is not providing enough information to appropriately infer that the true effect size is nonzero. And when the original study is very underpowered, where the influences of uncertainty and publication bias are stronger, even extremely large replication sample sizes may not be effective. Researchers are reminded of Hedges and Schauer's (2019a) warning: A single replication study is not always a panacea to an underpowered original study. These cases may also represent times in which the alternate strategy, such as ignoring the original study results and using other approaches for obtaining an appropriate sample size, may be more sensible.

Here we provide a more concrete situation. Suppose a researcher would like to replicate one of the findings of interest in a study on the effects of distracting objects in visual searches (Moher, 2020). In a 2 × 2 × 2 fully within-subjects ANOVA, the original researchers found a statistically significant two-way *target presence* × *distractor presence* interaction on error rates, $F(1, 199) = 54.14$, $p < .001$. The replicator can conduct sample size planning via the "Within ANOVA-General" BUCSS web app or the `ss.pwr.wa.general()` R function (see Figure 2 for the web app approach). After entering the observed *F* value (54.14), sample size (200), and numerator degrees of freedom (1) from the original study effect to be replicated, the replicator must enter the α level for the original study. This value represents the publication bias adjustment, or, the maximum *p*-value wherein the results would be available for the researcher to replicate. For published studies, entering .05 here is usually reasonable, given that the focal result may not have been published if $p > .05$, though this is in the control of the researcher. The α level for the replication study is also needed, for which our replicator selects .05. The final two inputs are the desired assurance and desired power. Suppose the replicator wants to achieve 80% power with a high degree of certainty, and chooses .95 for assurance. The full R code using the BUCSS package is:

```
ss.power.wa.general(F.observed = 54.14,
   N = 200, df.numerator = 1, alpha.prior = .05,
alpha.planned = .05, assurance = .95, power = .8)
```

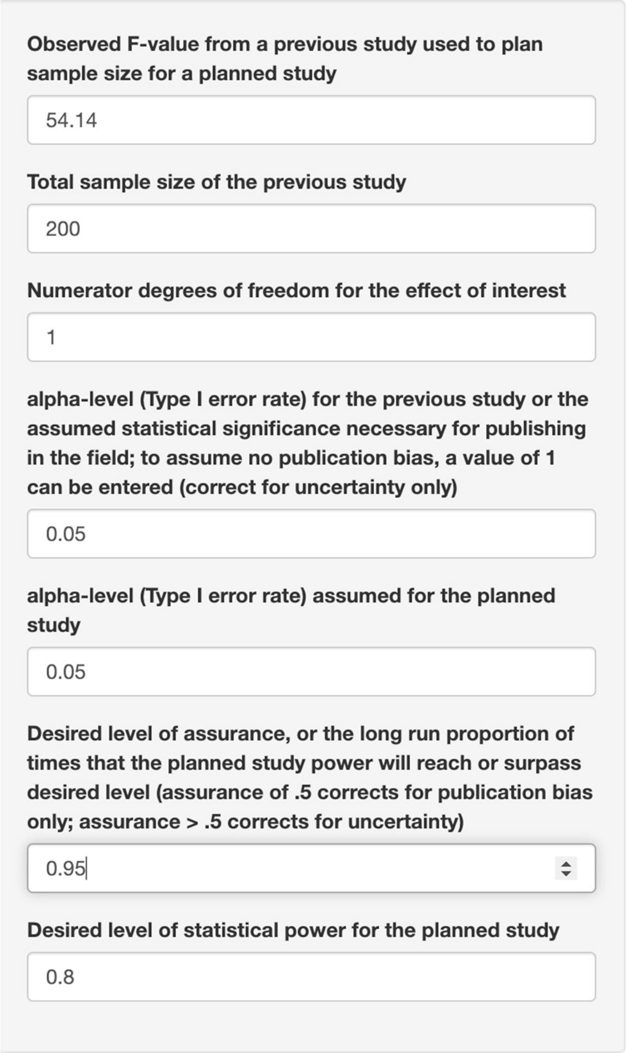
BUCSS thus suggests a sample size of $N = 53$ participants for the fully within-subjects replication study. Note that the sample size requirement here is not unreasonably large for most researchers in the empirical social sciences, even with such a high level of assurance, because the original study employed a powerful within-subjects design and demonstrated a large effect (even taking publication bias and uncertainty into account). In other cases, however, the replication study sample size may need to be much larger than the original.

Using NHST to define replication study successes and failures, and the accompanying sample size planning, may be appropriate if direction is much more of interest than size, if the statistical power of both studies can be very high, or perhaps among a suite of replication goals (see Discussion section). However, Schauer and Hedges (2021) showed that the proportions of false successes and false failures could often be above 20% using NHST, especially when the studies are underpowered or more unequal in terms of

[10] For example, in addition to its potential use in meta-analytic or accuracy-based methods, a reviewer noted that the BUCSS corrected noncentrality parameter could be used to plan the sample size for a replication study using Bayesian analyses.

Figure 2

Screenshot of a Researcher's Entries Into the BUCSS Web Application



Observed F-value from a previous study used to plan sample size for a planned study

54.14

Total sample size of the previous study

200

Numerator degrees of freedom for the effect of interest

1

alpha-level (Type I error rate) for the previous study or the assumed statistical significance necessary for publishing in the field; to assume no publication bias, a value of 1 can be entered (correct for uncertainty only)

0.05

alpha-level (Type I error rate) assumed for the planned study

0.05

Desired level of assurance, or the long run proportion of times that the planned study power will reach or surpass desired level (assurance of .5 corrects for publication bias only; assurance > .5 corrects for uncertainty)

0.95

Desired level of statistical power for the planned study

0.8

Note. BUCSS = bias-uncertainty corrected sample size.

power. Moreover, the NHST "vote-counting" approach to replication is not well-designed to accumulate results across a series of studies and can yield literatures that appear contradictory (Hedges & Olkin, 1980; Howard et al., 2000). Thus, researchers should place importance on careful sample size planning in both original and replication studies and consider the advantages offered by other perspectives and goals.

Before continuing on to describe sample size planning for alternative replication goals, it is important to return in more detail to the alternative way sample size can be conducted for Goal 1. As described, if the sample effect size from the original study is used in the sample size calculations for the replication study, this effect size should be adjusted for publication bias and uncertainty in order for the suggested sample size to result in adequate statistical power in the replication study. However, if there are concerns regarding the original study, particularly when the original study is quite underpowered, it may be difficult to adjust away all of the publication bias and sampling uncertainty or to ignore other issues. Moreover, even for direct replications, subtle differences in how variables are operationalized and tested (e.g., time of day) can have non-negligible effects on results (McShane & Bockenholt, 2014). In fact, even the very nature of the process being studied in the original study may have changed by the time the replication study is conducted (as an example to illustrate the point, consider what the attitudes toward remote-work were in 2019 compared with 2021). McShane and Bockenholt (2014) refer to this as between-study heterogeneity, meaning that the population parameters may not always be expected to be the same across replications. They showed that unaccounted for heterogeneity could lead to overly optimistic (i.e., too small) sample size suggestions, which is even more colored by the fact that publication bias and uncertainty were not considered.

In some cases, then, it may be sensible to ignore the original study and conduct the replication study with somewhat of a clean slate. While Goals 2–4 may also be relevant here, if obtaining a statistically significant replication effect is the goal, sample size planning can proceed using a theoretically relevant, population "minimally interesting effect size" as the effect size input instead of a sample value (that may be a poor approximation for the population value). The primary motivation here should be that there is an agreed upon value of the replication study effect size, under which is meaningless to detect. Standard sample size planning software readily conducts this type of sample size planning, and researchers expecting between-study heterogeneity could use McShane and Bockenholt's (2014) formulas to improve the procedure. We note that although the minimally interesting effect size approach can be used for Goal 1, to adequately power the replication study to detect an effect at least as large as the minimally interesting effect size, other goals, such as Goal 3, may be more appropriate in this case.

Goal 2: To Infer a "Null" Replication Effect

When replicating a study, as in any study, the existence of the purported effect is unknown. Yet, researchers' goals reasonably differ with respect to whether or not the original study effect is (qualitatively) believed to exist or not. When a replication study is developed in the NHST framework, the burden of proof rests on rejecting the null value (H_0), in which case a failure to do so would not represent convincing evidence that the original study's effect does not exist. This situation is reasonable when the researcher expects a true underlying effect, given that the replication study is set up to provide convincing evidence of that, at the cost of weak evidence in the event of a nonsignificant result. Despite this, replications yielding nonsignificant results are often interpreted as failed replications or null effects (Anderson & Maxwell, 2016). When researchers are surprised by a nonsignificant replication result, we discourage such overinterpretation. However, replication researchers are sometimes skeptical of original findings when designing replication studies. In these cases as well, nonsignificant findings using NHST replication (replication "failures") have been improperly used to suggest that the original study was a false positive or that the original authors engaged in *p*-hacking. A better strategy is to employ a method that is specifically designed to provide a more stringent test of the lack of an effect. Thus, when the motivation behind the replication is to refute an original finding, show that the effect is trivial, or suggest a false positive, the replication study can be designed to provide more convincing evidence of this, beyond simply failing to find a statistically significant result. Such an approach would be antithetical to what standard NHST can provide. Notably, the approach we subsequently describe can also be used in the event of a nonsignificant original study finding. Rather than directly comparing the original ambiguous findings to future replications, the replication researchers could plan a replication study specifically to appropriately demonstrate evidence of a trivial effect, replacing the original study with a study more rigorously designed for this purpose, in a sense.

In the frequentist framework, the two one-sided test (TOST) paradigm can be used to conduct an equivalence test, wherein the role of H_0 and the alternative hypothesis (H_a) are essentially flipped from what they would be for a traditional NHST (see Lakens et al., 2018. for a recent tutorial). In the Bayesian framework, the Region of Practical Equivalence (Kruschke, 2013) or Bayes Factors (Maxwell et al., 2015; Verhagen & Wagenmakers, 2014) can be used for similar purposes. Meta-analytic methods are another alternative when the goal is to show that an effect does not exist, focusing on, for example, a negligible meta-analytic effect. However, given the taxonomy here, this particular paradigm fits within Goal 4. In TOST specifically, researchers determine a priori the appropriate upper and lower bounds of an equivalence region, which is a set of values believed to constitute an essentially null effect. Said another way, this is an effect region that, at best, represents trivial effects that would not lead to action (e.g., claiming success, policy change, new protocol, etc.). The equivalence region, shown here on the standardized mean difference scale, $[\delta_L, \delta_U]$, is representative of the "good enough principle," which simply says that the effect is, at best, "good enough" to be considered a null effect (Serlin & Lapsley, 1985; see also Kelley & Maxwell, 2003). δ_L reflects the lower bound and δ_U reflects the upper bound, of the value of the population standardized mean difference, δ, deemed negligible. The hypotheses for δ can be written as:

$$\begin{aligned} H_0&: \frac{\mu_1 - \mu_2}{\sigma} \leq \delta_L \text{ or } \frac{\mu_1 - \mu_2}{\sigma} \geq \delta_U \\ H_a&: \delta_L < \frac{\mu_1 - \mu_2}{\sigma} < \delta_U \end{aligned} \tag{1}$$

where μ_1 and μ_2 are the population group means and σ is the common population standard deviation. Observed t-statistics can be calculated to test against each of the two sides of H_0 and compared to the appropriate $1 - \alpha$ critical t-value, hence conducting two one-tailed tests, both of which must be rejected to claim equivalence. Alternatively, a $100(1 - 2\alpha)\%$ CI for the effect can be calculated: equivalence is only claimed if this entire CI falls within the specified equivalence region.

Although equivalence testing has long been common in randomized controlled trials in the pharmaceutical industry (FDA, 2001), the practice has recently been gaining momentum in the social sciences but sample size planning has lagged behind (Guo et al., 2019). Yet, sample size planning for equivalence testing is critical, as much larger, even shockingly so, sample sizes are generally required than for traditional hypothesis tests. It is a strong claim to state that equivalence holds, and correspondingly the sample size will generally need to be larger than typically seen in psychology and related disciplines. The size of the equivalence region plays a major role in sample size and power. For example, when the equivalence region is wide in a particular context, relatively smaller sample sizes are required for a fixed value of the mean difference. Several writings and guides for equivalence testing assume a wide equivalence region (e.g., $-1 \leq \delta \leq 1$ or even $-2 \leq \delta \leq 2$), suggesting smaller sample sizes than described here (Guo et al., 2019; Shieh, 2016). However, it may not be fair to claim "equivalence" when the equivalence region is wide. Correspondingly, readers may be less convinced that an effect contained within a generous equivalence region really constitutes an essentially null effect. This is akin to claiming one's at-home cooking is not so different—essentially equivalent—to a Michelin three-star restaurant. A wide range of what is allowable for "essentially equivalent" may be a claim the at-home chef makes, but the wide range of allowable difference may make those tasting the food doubt the claim. Thus, one has to choose an equivalence range that is sufficiently narrow for the context, otherwise the claim will not be seen as valid. Unfortunately, perhaps, there is no one "correct" way to define the equivalence region.

Because the roles of H_0 and H_a are reversed compared with traditional NHST, power calculations for equivalence tests can be derived from central, rather than noncentral distributions (technical details can be found in Chow et al., 2008; Shieh, 2016; Zhang, 2003). Chow et al. (2008, Chapter 3 and 10) provide such methods for approximate sample sizes for various designs. We demonstrate the process for the case of an independent samples t test because (a) the conceptual foundations of equivalence testing are more easily understood for this design yet easily extended; (b) hypotheses regarding equivalence often naturally compare two groups; (c) sample size planning for equivalence is still relatively new and most software implementations are extremely limited outside of the two-group design (e.g., NCSS LLC, 2021); and (d) many important theoretical questions are addressed with two independent group designs. In the case of an independent samples t test, an approximate formula for the sample size per group is

$$n \approx \frac{2(z_{1-\alpha} + z_{1-\beta/2})^2}{\delta^2}, \tag{2}$$

where $z_{1-\alpha}$ is the standard normal variate at the $1 - \alpha$ quantile, $z_{1-\beta/2}$ is another standard normal variate at the $1 - \beta/2$ quantile, β is the type II error rate (= 1 – power), and δ is the value of the standardized mean difference quantifying what is essentially a null effect (i.e., the upper bound of the equivalence region). The formula uses $1 - \beta/2$ rather than $1 - \beta$ because two one-sided tests will be performed, each with a rejection probability of $\beta/2$, additively resulting in an overall rejection probability of β (Chow et al., 2008; Guo et al., 2019; Kieser & Hauschke, 1999). Equation 2 is technically an approximation, but works well as long as the true effect size is close to zero and the equivalence bounds are symmetric (Zhang, 2003). For many situations in psychology and related disciplines, it is reasonable to assume the true effect size is close to zero, given the skeptical view taken in Goal 2, but momentarily we show additional options for when the effect size is thought to be small, but nonzero (where Equation 2 can become somewhat conservative).

We now provide an example. Assume a replication researcher's goal is to replicate a study on priming (Bargh et al., 1996), where the original study found that students ($n_1 = n_2 = 15$; $N = 30$) primed with elderly related words walked statistically significantly more slowly out of the room than control subjects, $t(28) = 2.86$, $p < .01$, $d = 1.04$. Further suppose that this replicator is skeptical of the original study's findings—believing that the statistical significance might be a Type I error—and wants the planned replication to provide convincing evidence that the effect is, at best, minimal enough to be deemed null. Surprisingly, perhaps, no direct information from the original study is required to conduct sample size planning. Rather, the replicator must first select the upper and lower limits of the equivalence region: After consulting with the priming literature, the researcher decides that a value of δ between -0.15 and $+0.15$ is essentially null given all of the other influences on walking speed (i.e., $\delta_L = -0.15$, $\delta_U = 0.15$), meaning that even if the true value of the effect were contained within the interval $[-0.15, 0.15]$, the effect would be of no practical value. To some, this equivalence region may seem wide, but this is the beauty of such an approach: there is transparency and readers and reviewers are free to question such choices. The researcher can then solve for n in Equation 2, using $z_{1-\alpha} = 1.645$ and $z_{1-\beta/2} = 1.282$ for a standard .05 α level and 80% power, respectively, yielding:

$$n \approx \frac{2(1.645 + 1.282)^2}{0.15^2} = 761.5$$

We emphasize that (a) the sample size should be rounded up to 762 and (b) that this is per group. Thus, using this method, the replication will need 762 participants per group (1,524 total) to be powered appropriately to identify whether the replication study effect is small enough to be deemed trivial, emphasizing the large sample sizes needed unless the equivalence region is very wide.[11] Given the large sample size required, multiple replication studies may be needed to more convincingly demonstrate equivalence, a point we return to momentarily.

As previously discussed, Equation 2 is an approximation that will not be maximally accurate for all situations. Specifically, the approximation in Equation 2 will suggest sample sizes that are too large when the (unknown) population difference in means is nonzero but still might be considered negligible, especially if it is close to the boundary of the equivalence region. Zhang (2003) showed that the true power is close to .80 (i.e., $< .85$) when $\delta \leq 0.03$, but asymptotes at around .90 when $\delta \geq 0.17$. Around this asymptote, the true power is better approximated by the power of just one of the one-sided tests (Kieser & Hauschke, 1999). On the other hand, the alternative convention of using $1 - \beta$ results in sample sizes that are much too small when the population mean difference is closer to zero (i.e., underpowering by up to 20%; Zhang, 2003). A variety of imperfect solutions have been proposed, with some too conservative and others too liberal. In response, Zhang (2003) developed a unified formula that better approximates power for true mean differences closer to the boundary of the equivalence region, and, more recently, Guo et al. (2019) improved on this with an online calculator that employs an exhaustive local search algorithm for two-group designs. The calculator can be accessed from: https://optimal-sample-size.shinyapps.io/equivalence-of-means/. Importantly, when the details of our example above are entered into the calculator, the suggested sample size is identical to the approximation from Equation 2, but the calculator allows for heterogeneity of variance and unequal allocation, in addition to considering cost constraints. For some standard two-group designs, exact sample size formulations can also be implemented in R (see Shieh, 2016), although the approximations are reasonable most of the time. But we emphasize that with all of these approaches, including the exact formulations, the accuracy depends in part on how accurately the researcher specifies the unknown value of the true effect size, as with other approaches to sample size planning.

Regardless of the algorithm used, notice that the replicator's goal to demonstrate that the effect is, at best, "good enough" to be deemed essentially null requires a drastically larger sample size than the reverse of trying to demonstrate a nonzero effect. We provide a table showing equivalence testing sample sizes for various circumstances (see Table 2). In fact, the replication sample size (n = 762/group) is 50 times larger than the original study that was deemed publishable (n = 15/group). Ironically, this means that a researcher running original studies could run 50 studies using the same number of participants it takes one replicator to conduct a single study that refutes an original study's claim ([762 * 2]/[15 * 2] = 50.8). This emphasizes a cruel fact for replicators, which we term the Replicator's Dilemma, which is that when a study gets published, even one that is poorly conducted with respect to measurement, design, analysis, or p-hacking, it requires a disproportionate amount of effort to convincingly refute it. More generally, the Replicator's Dilemma could be colloquially described as a special case of Brandolini's Law (Uy, 2019). A stark example of the Replicator's Dilemma is the incredible amount of effort and resources have used to refute the fraudulent results of a small study that falsely claimed there was a link between the MMR vaccine and autism (see Rao & Andrade, 2011 for a timeline).

Even in a less unethical circumstance, if 50 separate studies were conducted involving a true null effect for the phenomenon, the expectation is that there would be around two or three Type I errors due to chance alone, and, the sheer number of studies will ensure a high probability (92%) of finding one or more statistically significant effects among these 50 studies.[12] A replicator may need to devote considerable time and resources simply to lay claim to the idea that the original study reports a false positive (i.e., that it does not replicate because the effect is not real). Further, as many publications do not value replications as much as they do novel findings, there is an extra burden placed on replicators whose intention is to correct a false claim in the literature. These dilemmas do not seem fair. The asymmetry in effort, which can be extreme, between the goal of simply trying to find something to publish (i.e., a statistically significant result) versus the goal of trying to correct the literature has not been appreciated by the methodological or applied research communities. We believe this has implications for how the scientific community awards prestige, usually where quantity of publications often is seen as more of a positive than quality. And, of course, things get even worse if multiple studies that found Type I errors are published. The Replicator's Dilemma thus plays an important role in the allocation of effort one may need to devote to a project.

[11] In R:

```
2*((qnorm(1-.05)+qnorm(1-.20/2))^2)/
  (.15^2).
```

[12] The approximation of 92% assumes that each of the 50 studies is independent and at least one finds a statistically significant effect. The following equation was used: $1 - (1 - .05)^{50}$.

Table 2
Approximate Sample Sizes for Equivalence Testing

Equivalence region	Statistical power	Per group *n*
$-0.25 < \delta < +0.25$	80%	275
$-0.25 < \delta < +0.25$	90%	346
$-0.20 < \delta < +0.20$	80%	429
$-0.20 < \delta < +0.20$	90%	542
$-0.10 < \delta < +0.10$	80%	1,713
$-0.10 < \delta < +0.10$	90%	2,165
$-0.05 < \delta < +0.05$	80%	6,852
$-0.05 < \delta < +0.05$	90%	8,659

Note. These sample sizes are calculated using Equation 2, which is approximate and can be somewhat conservative when the true effect size is not nearly zero. These calculations assume a test of two independent samples, normality, homogeneity of variance, and equal allocation to group.

Goal 3: To Precisely Estimate the Magnitude of the Replication Effect Size

Taking a more global perspective in building an informative, accurate, and cohesive scientific literature, the goal of replication, particularly systematic replication, is to build knowledge regarding a particular effect, with each replication contributing a piece to the puzzle. Goals 3 and 4 articulate methods that consider replication's role as a contributor to a "productive science" (Kelley & Maxwell, 2003, p. 319) framework. With Goal 3, researchers intend to move away from hypothesis testing and use the replication study to provide an appropriately accurate estimate of the magnitude of the effect size (measured via a narrow CI, in the classical paradigm, and a narrow highest density interval, in the Bayesian paradigm), which will generally be one more accurate than original study due to the fact that statistical significance is generally the key to publication in our current system and an accurate effect often, but not always, requires a larger sample size than a statistically significant effect. The current norms in psychology and related disciplines are that if sample size planning is formally conducted for original studies, it is usually conducted in terms of statistical power, rather than from an accuracy perspective, and often poorly so from even the power analysis framework.

Ironically, our experience with many grant proposals and discussions with researchers interested in submitting grants is that many studies (a) aim to estimate the magnitude of the effect size accurately but (b) plan sample size from a power analytic perspective. Thus, our experience is that there is often a mismatch between what is desired and how a study is planned. This mismatch is one of the issues that we hope this article solves, namely clearly delineating *how to plan replication studies based on the particular research goals*. Although reporting effect size and CIs are encouraged or required (e.g., Funder et al., 2014; Wilkinson, 1999), smaller sample sizes result in sample effect sizes upwardly biased to a greater degree (publication bias) and CIs that are "embarrassingly large" (Cohen, 1994, p. 1002). Some research has shown the replication study effect sizes tend to be smaller, a result of a system that often favors statistically significant findings for original studies (see Held et al., 2020 for an interesting perspective incorporating publication bias and regression to the mean), although a recent review of replication projects found mixed evidence for this (Schauer & Hedges, 2020). The motivation to consider Goal 3 for replication could be based on a belief that the original study effect size is spuriously large or simply a desire to set up the study with accurate parameter estimation in mind, rather than reporting effect size and CIs as an afterthought. The more accurate effect size estimate in the replication enhances the literature by yielding more certainty regarding the magnitude of the purported effect.

When the goal is estimating magnitude, sample size planning should be conducted using accuracy-based procedures. Note that although these procedures can require a larger sample size than power analysis would suggest, this is not always the case, especially when the population effect size is small (Kelley & Maxwell, 2003). In other words, power and accuracy are two separate goals. A particular paradigm that has been developed and extended to a variety of designs using classical methods is accuracy in parameter estimation (AIPE; technical details described in Kelley, 2007; Kelley & Maxwell, 2003; Kelley & Rausch, 2006), aimed at calculating the sample size that will optimize the width of a CI for the population effect to a researcher's desired level of accuracy. In other words, rather than focusing on whether the CI contains the null value or not (hypothesis testing), the idea is to decrease the size of the range of plausible values for the effect (i.e., make the CI narrower; Kelley & Rausch, 2006), improving accuracy by reducing bias and increasing precision. Mathematically, accuracy is the square root of the mean square error, $\sqrt{RMSE}$, which decreases as CI width decreases, when coverage is held constant (Kelley & Rausch, 2006).

In addition to achieving an *expected* appropriately accurate estimate of effect size on average, researchers may want to go above and beyond to also incorporate a *degree of assurance* that the observed CI will be sufficiently narrow. For many types of effects, AIPE has been developed to allow the researcher to have a desired assurance that the observed computed interval will be sufficiently narrow (Hahn & Meeker, 1991; Kelley & Rausch, 2006). This is a direct parallel to the assurance concept described under Goal 1: *AIPE can provide both desired accuracy and assurance that the desired accuracy can be achieved.* The mathematics become somewhat more complex, in general, but the process is straightforward with existing software (e.g., via the MBESS R package, Kelley, 2007) or with a priori Monte Carlo simulations (e.g., Maxwell et al., 2008). MBESS contains functions for conducting AIPE for a variety of parameters within standard and more complex designs, including the standardized mean difference, ANOVA/ANCOVA contrasts, R^2 in a multiple regression context, unstandardized and standardized regression coefficients, reliability metrics, RMSEA and SEM path coefficients (Lai & Kelley, 2011), and parameters in cluster randomized trials (Pornprasertmanit & Schneider, 2014).

Even if researchers decide they want to plan sample size from an AIPE perspective, one problem still remains. The most difficult aspect to implement is specifying an appropriate value of the population effect size to use. This is analogous to the way in which power analysis requires one to specify one or more population parameters, though AIPE sometimes requires fewer parameters to specify. That said, AIPE formulas do not completely circumvent the need to specify one or more effect sizes. There are some suggestions to overcome this limitation. For example, in a multiple regression context, Kelley and Maxwell (2003) suggest simplifying assumptions (e.g., exchangeable correlation structure among the predictors), hypothetical population values, or values reported in prior research (particularly meta-analytic values, if available). They also warn against directly relying on values estimated from pilot studies due to the potential for large sampling error and encourage a sensitivity analysis to assess how robust the required sample size is to input parameters. Yet, the width of the CI still depends on the population effect size in order for the desired properties of the procedure to hold: This dependency varies depending on the parameter in question. In some cases, the impact of a small misspecification (e.g., using an uncertain or biased sample effect size estimate) can be large, while in others, the impact is little to none in practical situations, depending on the parameter of interest and sometimes the magnitude of the population effect (Kelley et al., 2018). We begin by focusing on the latter case. Consider the standardized mean difference, δ, and the desire to determine the sample size needed to estimate it accurately. We show how AIPE applies in this situation because of how common δ is as a primary outcome of interest (e.g., Chattopadhyay & Kelley, 2017).

Suppose a researcher is attempting to replicate a study finding mental benefits of ballroom dancing in older adults with dementia (Lazarou et al., 2017). The original study ($N = 129$) reported means and standard deviations for the intervention ($M = 25.65$, $SD = 3.27$) and control groups ($M = 28$, $SD = 2.39$) postintervention, and Cohen's d can be easily calculated from the reported information: $d = 0.82$ for the effect of intervention on the Mental Mini Status Exam. The researcher wants to design a study to better estimate the magnitude of the effectiveness of the dance intervention, to aid in determining whether resources should be allocated to setting up or expanding such a program. The researcher decides that the level of accuracy desired is represented by a CI with a half-width no larger than 0.2 (full width 0.4), and that the desired assurance is 80%. Using the MBESS R function `ss.aipe.smd()`, the researcher can enter the following code:

```
ss.aipe.smd(delta = .82, conf.level = 0.95, width =
   .4, which.width = "Full", assurance = 0.8)
```

The researcher will need $n = 212$ participants per group to achieve the desired CI width with the desired assurance. If the researcher was concerned about uncertainty from using a sample estimate for δ, a sensitivity analysis could be conducted. However, consistent with what we noted in the previous paragraph, the necessary sample size is minimally impacted: If δ is as small as 0.1, $n = 193$ and if δ is as large as 1.0, $n = 221$.

In other cases, using a biased or uncertain point estimate from the literature in place of the population value is more impactful, such as for unstandardized contrasts (e.g., in an ANOVA or ANCOVA framework). For unstandardized contrasts, the size of the effect is independent of the effect's accuracy. Due to the independence of the mean and the variance of a normal distribution, the center of a CI has nothing to do with its width. The variance, though, very much impacts the CI width. Correspondingly, if the variance from a study is less than the population value and that value is used to plan AIPE for an unstandardized contrast, the sample size estimate will tend to be an underestimate the optimal sample size needed to accomplish the specified goals. Importantly and not often considered, one of the two primary ways in which publication bias presents itself is with an overly optimistic (i.e., too small) estimate of the population variance. Even if the unstandardized contrast itself is exact, a variance that is too small will lead to a t-statistic that is unduly large due to the role the standard deviation plays in the denominator of t-tests. Correspondingly, the p-value will be smaller than it otherwise would be (e.g., if the values more closely approximated the true values in the population) and thus more likely to achieve statistical significance (and thereby be publishable under a publication model that requires statistically significant results for publication).

Thus, in some situations, specifying the input parameters (e.g., the mean difference) is very important, yet in other situations it is less so (e.g., δ). Because of this difficulty, Kelley et al. (2018) developed a solution that employs AIPE within a sequential estimation paradigm (a technical presentation can be found in Chattopadhyay & Kelley, 2017; Kelley et al., 2018, 2019). Sequential methods are a developing set of procedures for obtaining accurate estimates that allow sample sizes to be increased as needed throughout the study, such that data collection stops only when a prespecified stopping rule is reached. In the NHST framework, repeated analysis of the data—as more data are collected and the analysis is redone—increases the Type I error rate, and that is thus an inappropriate stopping rule. When necessary, such as to evaluate if a clinical trial should be stopped, this is handled by α spending procedures, where, for example, the α level can be allocated to several intermittent tests, each spending a portion of the total amount (DeMets & Lan, 1994). The sequential version of AIPE, though, because it is unconcerned with the center of the CI or what the CI may or may not contain (e.g., zero), continues sampling until the CI is as narrow as needed: Because the decision to continue collecting data is not based on statistical significance (i.e., whether the CI does or does not include the null value), the procedure evades this complication. Moreover, the procedure as implemented in Kelley et al. (2018) does not require specifying any value for the effect size in advance and is from a distribution-free framework and therefore does not require normality. The same logic applies for those that would rather make parametric assumptions. Here, we highlight the CI formulation for sequential AIPE as discussed in Kelley et al. (2018) which is:

$$CI = \left(T_n - z_{1-\alpha/2}\sqrt{\frac{\xi^2}{n}},\ T_n + z_{1-\alpha/2}\sqrt{\frac{\xi^2}{n}}\right), \tag{3}$$

where T_n is the parameter for which the CI will be formed (e.g., δ, or an unstandardized regression coefficient), $z_{1-\alpha/2}$ is a $1 - \alpha/2$ critical value from a standard normal distribution, and ξ^2/n is the asymptotic variance of T_n. In general, sequential AIPE ensures that the resulting sample size yields an interval no larger than width ω. Notice that the z-distribution (not the t-distribution) is used, which illustrates that this procedure is best used for sample sizes that are not too small. That said, as we have discussed, for realistic situations the sample size will generally not be "too small" for an AIPE approach to sample size planning to use a z critical value instead of a t critical value.

The procedure is conducted as follows, using generalized notation, allowing the method to be useful in any situation in which the Central Limit Theorem applies. First, the researcher chooses the desired CI width, ω. Second, a pilot sample size, m, is chosen based on the larger of two alternatives: (a) m_0, the smallest sample size needed to estimate ξ^2 or (b) the smallest integer not less than $2z_{1-\alpha/2}/\omega$. The researcher uses this sample size, m, to estimate ξ^2.

Then, using this estimate, $\hat{\xi}^2$, the researcher then checks the following stopping condition:

$$m \geq \left[\frac{4z^2_{1-\alpha/2}}{\omega} \left(\hat{\xi}^2 + \frac{1}{m} \right) \right]. \tag{4}$$

In the event that Equation 4 is satisfied, the researcher stops data collection at the initial pilot sample. If not, the researcher continues to the second step, collecting an additional m' participants. Choice of m' may be difficult, but can be decided upon based on cost and feasibility. For example, if it is almost as easy to collect an additional $m' = 10$ participants than collecting a single participant, the larger m' can be used. With the new sample size, $m + m'$, the researcher reestimates ξ^2 and again checks the stopping condition, replacing m in Equation 4 with $m + m'$. If Equation 4 is not satisfied, the researcher continues to repeat the procedure. The final per-group sample size will be that which satisfies Equation 4.

There are potential limitations to sequential AIPE: Its flexibility increases the uncertainty, as the final sample size is not known in the beginning of the study. but the general framework solves the issue of the "problematic parameter" (Lipsey & Aiken, 1990, p. 47). Readers interested in the technical details of the procedure are referred to Kelley et al. (2018), but an accessible, nontechnical demonstration of using the method in practice is detailed below. Importantly, a detailed understanding of Equations 3 and 4 is not necessary to employ sequential AIPE in practice, as software handles most of the heavy lifting.

Suppose a researcher is attempting to replicate a study assessing the relationship between years since obtaining a PhD and academic-year salary in a particular field. Because sequential AIPE requires sample size analysis during the study itself, the example will suppose that the full data set is not yet already collected. The Salaries dataset (`CarData` package; variables `years.since.PhD` and `salary`; Fox & Weisberg, 2019) will be used as the hypothetical researcher's replication data. Of course, in reality, the full data set would not be available, but sampling would continue as data comes in until the stopping rule is satisfied. The necessary functions for sequential AIPE are available in the Sequential Methods for Study Design (SMSD; Kelley, 2020) R package.[13]

In the initial step, suppose the researcher selects $\omega = 500$, based on knowledge of the scales of the variables being studied. The goal is thus to obtain an interval whose bounds are within $ ± 250 dollars of the average slope, so as to have what is thought to be a sufficiently accurate estimate of this primary parameter of interest. The researcher first calculates the initial (pilot) sample size m using the following code:

```
seq_aipe_slr_slope(alpha = .05, omega = 500,
   pilot=TRUE)
```

The researcher specifies the desired alpha level (i.e., Type I error rate; .05), the desired CI width (500), and notes that a pilot sample size is desired (`pilot = TRUE`). The function suggests $m = 4$. Thus, the researcher collects data from $m = 4$ participants (here, "collected" as the first four participants from the Salaries data). Now that the researcher has collected a pilot sample, the `seq_aipe_slr_slope` function can be used to determine whether this initial sample size is sufficient or not. The inputs are as follows:

```
seq_aipe_slr_slope(alpha = .05, omega = 500,
 data=data.pilot)
```

The researcher specifies the desired alpha level (.05) and CI width (500) as before, as well as the name of the dataset containing pilot data collected so far, (`data = data.pilot`). The function provides four pieces of output: (a) the current sample size, (b) the current estimate of the slope, (c) whether the stopping criterion is satisfied, and, if the stopping criterion is satisfied, (d) the current CI. Not surprisingly, with the pilot data, the stopping rule is not satisfied. The researcher would next need to decide how many additional participants to collect, and decides to check the stopping criterion after every $m' = 17$ additional participants. Note that the procedure assumes m' is a fixed quantity (i.e., the same number of participants are collected between checks of the stopping criterion), but in practice, m' does not necessarily need to be constant. However, with larger or uneven choices of m', the researcher can run the risk of oversampling (e.g., 101 participants needed, but 120 participants collected). The above code is repeated, replacing (`data = data.pilot`) with the name of the new cumulative dataset until the stopping criterion is met. In this scenario, the stopping criterion is met after collecting $N = 395$ participants through 23 rounds:

```
seq_aipe_slr_slope(alpha = .05, omega = 500,
    data=data.395)
   [1] "The stopping rule has been met; sample
    size is large enough."
   $Current.n
   [1] 395
   $Current.slope
   [1] 982.1311
   $`Is.Satisfied?`
   [1] TRUE
   $`Confidence Interval`
   [1] 733.5191 1230.7431
```

The salary researcher can now discontinue data collection, with a CI width as narrow as desired. Kelley and Rausch (2006) note that AIPE procedures can also be used to plan sample sizes for equivalence testing, when the CI formulation is used. Expanding on their idea, the motivation would be to plan the appropriate sample size for a 90% CI to be sufficiently narrow so that it could fit completely within the equivalence region, should the center of the CI be close to zero.

The sequential AIPE approach offers great flexibility in terms of not needing to specify unknown population effect size values and stopping data collection when the sufficient accuracy is achieved. Such an approach, though not yet used frequently in practice, will provide researchers with a flexible approach to advance their research agendas. Previous treatments of such a sequential sample size planning procedure for narrow confidence were overly technical and did not provide software that was easy

[13] SMSD can be installed via:

```
devtools::install_github("yelleKneK/SMSD").
```

to use, so we have purposely provided a more accessible treatment here that we believe will be useful to researchers.

Goal 4: To Combine Results Across Original and Multiple Replication Studies

The scientific literature is supposed to be trustworthy. Yet, researchers have demonstrated that these ideals are not always achieved in practice when making inductive decisions based on a single original study (e.g., Anderson, 2020) or even when pairing an original study with a single replication study, due in part to concerns about statistical power (Maxwell et al., 2015; Schauer & Hedges, 2020). Taking a broader perspective on the cumulative knowledge gained across studies in an area of interest may be particularly helpful, as Cumming (2014) noted, "any one study is most likely contributing rather than determining" knowledge (p. 23). This in fact harkens back to Fisher, who considered statistical significance to be decisive only if repeatedly found over multiple experiments (see Goodman, 2016). From the meta-analytic perspective, replication can be seen as a process that continues over time as multiple studies are conducted sequentially (or in parallel), rather than a procedure that is conducted only once (e.g., Braver et al., 2014; Shrout & Rodgers, 2018). Thus, another replication goal is to combine information from an original and replication study, with the more distal goal of conducting multiple replication studies and concatenating the knowledge into a more unified and trustworthy conclusion. We focus on meta-analysis from the classical perspective, which has received the bulk of the research, but readers can consult other sources for an introduction to Bayesian meta-analysis (Kruschke & Liddell, 2018), a meta-analytic Bayes Factor (Rouder & Morey, 2012; Verhagen & Wagenmakers, 2014), and a Bayesian meta-analytic approach that directly incorporates replication (Carter & McCullough, 2018).

Jointly aggregating multiple studies, with the hope that bias is not increasing across such studies, improves both statistical power and the precision of the effect size estimate (Bonett, 2009; Howard et al., 2000), as opposed to individually assessing the potentially more contradictory literature involving underpowered NHST replication attempts. In fact, a nonsignificant replication when viewed alone may result in a statistically significant effect when combined with an original effect size, and maybe even yield a smaller *p*-value than the original study (Anderson & Maxwell, 2016; Braver et al., 2014), in addition to providing a narrower CI. To appropriately summarize data from groups of studies, meta-analytic methods can be implemented. To illustrate the strengths of meta-analysis, Howard et al. (2000) showed a clever example where a meta-analysis including two nonsignificant findings and one significant finding rejected the null hypothesis. Readers can consult Borenstein et al. (2009) and Hunter and Schmidt (2004) for thorough guides to conducting meta-analysis.

Meta-analytic approaches can be geared toward either estimating a population parameter across multiple studies or on the degree of between-study heterogeneity in the effect estimate (Shrout & Rodgers, 2018). Both of these foci may be of interest to a replicator. Estimating a parameter is applicable when the desire is to corroborate the existence of the overall effect or estimate its magnitude with sufficient accuracy, whereas the degree of between-study heterogeneity is applicable when the desire is to test or compute a CI for the difference/variation in effect sizes across studies. Goal 4 is specifically concerned with using meta-analysis to estimate one or more parameter values.[14] Notably, Schauer and Hedges (2021) consider this "an analysis assuming a successful replication" (p. 11), given that (a) the success or failure of individual replication attempts is not quantified, and (b) this goal expects that the underlying effect is the same or similar across replication studies.

There are a variety of choices involved when averaging effect sizes across studies. One is whether to employ a fixed- or random-effects model. Fixed-effect methods have been historically common and, as we will describe, can yield better precision than random-effect methods (Bonett, 2009). However, fixed-effects methods assume that the population effect size is identical across studies and estimates can incur bias if this is violated (Bonett, 2021). Random-effects models, on the other hand, do not assume a constant effect size (in the spirit of McShane & Bockenholt, 2014), but assume that the studies represent a random sample from a (typically normal) superpopulation of studies. That the selected studies are truly a random sample is potentially violated, and these and related assumptions can be more challenging to assess with a small number of replication studies (Bonett, 2009, 2021; Schulze, 2004). Additionally, due to the inclusion of between-study variation, random-effects models provide a remarkably imprecise estimate of the combined effect with few studies, even when the within-study sample sizes are large (Borenstein et al., 2010; Verhagen & Wagenmakers, 2014), making this a poor choice for a single replication study. More recently, a new class of fixed-effects models has been proposed, the varying-coefficient method, which bypasses both the random sampling and constant effect size assumptions. Varying-coefficient formulas have been derived for coefficient alpha (Bonett, 2010), odds ratios and risk differences (Bonett & Price, 2015), correlations (Bonett, 2008), and standardized mean differences (Bonett, 2009). The latter two implementations do not assume homogeneity of variance within and across studies.

The questions of relevance here are which methods are appropriate for the case of replication studies and whether these distinctions can presently be applied to sample size planning. In reference to replication, Schauer and Hedges (2021) noted that the fixed-effects model can be appropriate when a common underlying effect size is assumed, which is the situation in which Goal 4 has the most conceptual meaning, and Borenstein et al. (2010) take this position more generally (though, see Bonett, 2021). In the case of highly controlled direct replication studies, or when an underlying effect is thought to be robust to minor differences in settings, we contend that this assumption may be reasonable. In reference to sample size planning from the perspective of a replicator planning a new study as opposed to a researcher conducting a meta-analysis from available studies, there is not yet a clear formulation when using varying-coefficient methods for this purpose.[15] Given these and other complexities, we contend that the sample size planned from a fixed-effects approach can still be useful, even when a random-effects or varying-coefficient meta-analysis is ultimately used to analyze the data, and would serve as

[14] A reviewer aptly pointed out that this interpretation of Goal 4 could be considered "Goal 3b". To align with previous work and the varying sample size planning methods that are used, we keep these goals separate, but note that they can indeed be seen as related.

[15] Bonett (2021) provides R functions for conducting varying coefficient meta-analysis, but sample size functions have not yet been developed for these methods.

a lower-bound sample size (see also Raudenbush & Liu, 2000). The example we show below relies on fixed-effects methods, but we recommend further development of varying-coefficient models, and sample size planning guidance in particular, given that variability among effects can be surprisingly significant even in controlled settings (Klein et al., 2018).[16]

The meta-analytic literature has typically been concerned with a fixed number of already conducted studies, and thus avoids the topic of sample size planning for a new study (see Valentine et al., 2010 for a detailed primer on statistical power considerations for meta-analysis). Yet as a replication goal, researchers can explicitly consider sample size prior to conducting the replication study, which will then be meta-analyzed alongside the original study and any prior replication attempts. The following paragraphs organize equations that can be used to conduct sample size planning for a new study to be included in a meta-analytic replication, a process then demonstrated with an example. Consider the case of an independent samples t test. In fixed-effects meta-analysis, a weighted average, $\overline{d}$, of the effect size estimate can be calculated from

$$\overline{d} = \frac{\sum_{i=1}^{m} \frac{1}{v_i} d_i}{\sum_{i=1}^{m} \frac{1}{v_i}}, \tag{5}$$

where d_i is the standardized mean difference from study i $(i = 1, \ldots m)$ and $1/v_i$ is an individual study weight, calculated based on the study's precision, or the inverse of the within-study variance for study i, v_i (Cohn & Becker, 2003). The within-study variance of d is

$$v_i = \frac{n_{1i} + n_{2i}}{n_{1i} n_{2i}} + \frac{\delta_i^2}{2(n_{1i} + n_{2i})}, \tag{6}$$

where n_{1i} and n_{2i} are the sample sizes from Group 1 and 2 in Study i, Δ and is often estimated with d_i as defined above (Hedges, 1982). As referenced in Cohn and Becker (2003), the variance of $\overline{d}$ is approximately

$$\operatorname{var}(\overline{d}) \approx \frac{1}{\sum_{i=1}^{m} \frac{1}{v_i}}. \tag{7}$$

Thus, a CI for the meta-analytic average of d across m studies is given by,

$$\overline{d} \pm z_{1-\alpha/2} \sqrt{\operatorname{var}(\overline{d})}, \tag{8}$$

where $z_{1-\alpha/2}$ is a standard normal critical value at $1 - \alpha/2$.[17] As the total sample size and total number of studies involved in the meta-analysis increases, the width of the CI decreases (Cohn & Becker, 2003; Liu, 2015). In terms of statistical power (i.e., testing whether the meta-analytic average differs from zero), the relevant noncentrality parameter is

$$\lambda = \frac{\delta}{\sqrt{\operatorname{var}(\overline{d})}}, \tag{9}$$

which shows that power to detect the effect increases as the population effect deviates from zero and as the variance across studies decreases. This should come as no surprise, as this principle holds across statistical power procedures generally, holding other factors constant.

In meta-analysis, the researcher may have to consider both the number of studies, m, and the number of participants per study, N_i, as sample sizes (Valentine et al., 2010). A researcher who is planning to conduct a meta-analysis based on a set of published studies may need to consider whether the desired power or accuracy for the overall effect can be achieved, given fixed per-study sample sizes and the number of studies available. Building on what we described previously, the perspective we take is that from a series of replication studies, wherein the researcher wants to calculate the sample size needed in the m^{th} study, to achieve the desired accuracy when averaged with previous studies. Thus, m is considered fixed. For simplicity, the sample size planning example that follows assumes the researcher is conducting the second replication study on a topic, thus $m = 3$, and that the planned replication study (independent t test) will attain equal per-group sample sizes.

We now explain the process of sample size planning for Goal 4 in general terms before illustrating with real data. To begin, the researcher first determines the desired half-width of the CI for the overall population δ. The researcher can use the right side of Equation 8 to determine the value of $\operatorname{var}(\overline{d})$ that results in the desired half width. Second, the variance of d in each prior study can be calculated, using Equation 6. Third, the researcher can use Equation 7, setting it equal to the necessary value of $\operatorname{var}(\overline{d})$ determined in the second step, solving for the variance of the m^{th} study. Fourth, now that the desired variance of d for the replication study has been calculated, the researcher can use Equation 6 to solve for the per-group sample size that achieves this variance.

Because the value of d needed to estimate sample size using Equation 6 has not been calculated, as the study has not been conducted, the researcher could input the meta-analytic $\overline{d}$ estimate from previous studies, adjusting for estimation bias using a formula appropriate for the parameter of interest. For the case of the independent t test, the relevant bias adjustment is

$$b_i = 1 - \frac{3}{4(n_{i1} + n_{i2}) - 9}, \tag{10}$$

where n_{1i} and n_{2i} again refer to the sample sizes of Group 1 and 2, respectively, in study i. It is well-known that the usual estimate of Cohen's d is a biased estimator of δ, but the level of bias is typically negligible unless the sample size is extremely small (e.g., < 25). However, given varying estimators of δ and varying sample sizes across studies, Bonett (2009) recommends applying this bias adjustment to each effect size, which can be done using software such as MBESS (e.g., `smd` function. Though there will often be little benefit, there is no harm in doing so.

There is one more caveat to entering an estimated parameter value in the fourth step in the process. Meta-analysis reduces

[16] Readers can also reference Valentine et al. (2010) for relevant power analytic formulas relevant to the random effects case.

[17] Underlying normality is assumed, but this CI does not become more robust to non-normality as sample size increases. Bonett (2009) noted that this assumption is often reasonable in psychology, but should be taken seriously when calculating CIs for standardized mean differences within and across studies. Bonett's (2009) varying coefficient method also assumes normality.

uncertainty in the overall estimate of the parameter in question but does not by default adjust for publication bias. Notably, adjustments for publication bias will typically increase the variance of the estimates, reducing meta-analytic statistical power and accuracy. Publication bias adjustment can be accomplished using methods such as BUCSS (see Goal 1), Hedges' (1984), or McShane et al. (2016), among others.

So as to put the preceding steps into practice, which have not been well-documented within the context of planning a new replication study in the extant literature, we provide an example. Suppose that a researcher has seen the first two published studies reported in Table 1 of Cohn and Becker (2003), investigating gender differences in risky driving behaviors. The original study reports $d = 0.03$, with 28 and 40 participants in the two groups, respectively. The first replication study reports $d = 0.53$ with 71 and 118 participants, and the meta-analytic $\overline{d} = 0.39$. For simplicity, assume that the reported effect sizes have already been adjusted for both estimation bias and publication bias. First, the researcher determines that a half-width of 0.2 represents the desired level of accuracy of the meta-analytic average of the two previous studies and the planned replication study, such that if the meta-analytic $\overline{d}$ continues to be 0.39, the 95% CI would be [0.19, 0.59]. Using Equation 8, the corresponding $\text{var}(\overline{d})$ needed is 0.0104; based on solving Equation 8 with the following terms:

$$0.2 = 1.96\sqrt{\text{var}(\overline{d})}.$$

Second, using Equation 6, the researcher calculates that the variance of d is 0.0607 and 0.0233 in the original and first replication studies, respectively.

$$v_1 = \frac{28 + 40}{28 \times 40} + \frac{0.03^2}{2(28 + 40)} = 0.0607$$
$$v_2 = \frac{71 + 118}{71 \times 118} + \frac{0.53^2}{2(71 + 118)} = 0.0233$$

Third, solving for the necessary variance of the planned replication study in Equation 7:

$$0.0104 = \frac{1}{\frac{1}{0.0607} + \frac{1}{0.0233} + \frac{1}{v_3}}$$
$$v_3 = 0.0272$$

Now, the researcher can solve for the necessary sample size for Study 3 by reorganizing Equation 6, assuming equal n in the planned study,

$$0.0272 = \frac{2n}{n^2} + \frac{0.39^2}{4n}$$
$$n = 74.9$$

The researcher thus needs 75 participants per group—assuming equal sample sizes per group—in the planned replication study for a desired CI half-width for the average effect size to equal 0.2. An advantage of this approach is that the planned study sample size directly benefits from the data collected in the previous studies, and we expect this goal to become more common as multiple replications become more commonplace.

Bayesian Sample Size Planning for Replication

In Goals 1–4, we used a classical approach, but cited a variety of Bayesian manifestations of each goal. When using such analyses, the corresponding sample size planning procedure can be Bayesian. Bayesian analysis is well-suited for replication, founded on the principle of research as a process to provide information, weighting new studies by considering both prior knowledge, and new knowledge from the data at hand. Further, it formalizes the idea that parameters have a distribution (i.e., are not fixed). With the frequentist approach, incorporating uncertainty can sometimes be forced, whereas with the Bayesian approach, uncertainty is part of the framework itself in terms of the posterior distribution, which we say more about momentarily. An overview of Bayesian statistics is beyond the scope of the present article, but readers can consult Kruschke (2014) and Gelman et al. (2013) for more applied and technical introductions, respectively.

Two fundamental aspects of Bayesian statistics for understanding sample size planning and replication are the prior and posterior distribution. The prior distribution reflects plausible values, with differing probabilities, for an effect size from previous knowledge (i.e., a previous study). The analysis of a new study results in a posterior distribution, which reflects plausible values with differing probabilities, based on the combination of the prior and new data. The process of sample size planning in a Bayesian context is a general procedure that can be tailored to specific (replication) goals. Regardless of what the researcher's goal is or how the goal will be evaluated, Bayesian sample size planning follows a similar series of steps, using Monte Carlo simulation. Sample size planning should be tied to some measurable outcome, but there is flexibility about what that outcome can be. Notably, the literature on Bayesian sample size planning is still nascent, particularly with respect to Bayes Factors (Schönbrodt & Wagenmakers, 2018), and has not yet been clearly optimized for the context of most replication goals. However, the procedures are readily customizable to various replication study hypotheses and the paradigm is flexible. The following paragraphs outline the most general procedure, which is in the spirit of Maxwell et al.'s (2008) quote: "sample size can be planned for any research goal, on any statistical technique, in any situation with an a priori Monte Carlo simulation study" (p. 553).

First (Step 1), the researcher hypothesizes a probability distribution based on idealized data that considers the expected (unknown) effect in the planned study. This probability distribution can range from diffuse (acknowledging that much is unknown) to precise (where much knowledge already exists). For context, frequentist power analysis falls at the precise end of this continuum, assuming by default that the necessary values are specified without error. Uncertainty adjustments, such as sensitivity analyses or the assurance concept invoked in Goal 1 can somewhat approximate the more diffuse end of the continuum, but cannot assign probabilities to these effect sizes. A more Bayesian definition of assurance is the "probability of achieving a researcher's goal, averaged across all possible effect sizes" (Schönbrodt & Wagenmakers, 2018,

p. 131). This necessitates determining prior probabilities for each effect size, calculating an outcome distribution for each effect size, and computing a weighted average.

Before describing the remaining steps in this general paradigm for Bayesian sample size planning, we provide more context for how the idealized data probability distribution is defined using prior probabilities. Schönbrodt and Wagenmakers (2018) importantly distinguish design priors, which are formed for necessary parameters during sample size planning, from analysis priors. A criticism of Bayesian statistics is presumed subjectivity and overreliance on belief. Yet, in the design phase, previous knowledge becomes particularly important in optimally determining sample size, especially for replications. A more informative design prior, which at the most extreme specifies no distribution for the population effect size (i.e., a single value), can be somewhat risky, providing the most efficient (i.e., optimal) sample size if the researcher is perfectly correct, but leading to an underpowered study if the researcher is wrong in one direction and overpowered if wrong in the other direction. A more diffuse design prior will typically suggest larger sample sizes, which may be inefficient on the one hand, though justified if there is limited prior knowledge. Just as when specifying the desired assurance within the BUCSS procedure, AIPE, or when using sequential methods, researchers conducting Bayesian sample size planning should take a balanced approach to utilizing prior information and consider the tradeoff between informativeness and efficiency (Dupont & Plummer, 1990).

We now outline the remaining steps of this general Bayesian sample size planning procedure. In Step 2, the researcher draws a random sample from the initial idealized distribution derived in Step 1, which results in a single simulated dataset that would be plausible given the initial assumptions the researcher made in Step 1. The size of this sample can be the sample size the researcher reasonably expects may be adequate for their planned study, though if this initial guess is an underestimate, the process can be repeated with a larger sample size. In Step 3, the intended analysis is conducted on this simulated dataset, and the resultant metric of interest is calculated (e.g., Bayes Factor, posterior distribution). In Step 4, this process (Steps 2 and 3) is repeated many times (e.g., 10,000) to repeatedly generate representative simulated data sets and conduct the proposed Bayesian analysis on those data sets. The researcher can then examine the distribution of all replicates to infer the proportion of times the goal was met using the selected sample size. Finally, the researcher can increment this sample size up until the proportion of replicates that achieves the goal equals or exceeds the desired level of power. A posterior distribution for power can be formed from the proportions of times where the sample size used achieved the goal (beta distribution assuming a uniform prior for power), which becomes more precise as the number of Markov Chain Monte Carlo samples increases. For all practical purposes, a single value of power can be determined.

This general process can be amended in different ways in the replication context, and when different goals are of interest. However, it is important to note that Bayesian analysis is not an invitation to use small samples. Although small sample Bayesian methods exist for very limited sample sizes (McNeish, 2016), small samples will typically either yield inconclusive results (under less informative priors) or place little weight on the new data (under more informative priors).

Discussion

As the value of replications in psychology and related fields has increased, while also becoming more popular, researchers who attempt to replicate studies have tended to use the existing methods for original studies to design and analyze replication studies. However, considering alternative goals for replication studies, which are different than original studies, can increase the effectiveness of the planned replications and thereby improve the literature by making it a more cumulative source of knowledge with more accurate conclusions. The literature has recently extolled the virtues of a variety of different goals and associated analyses for replication, but sample size planning specific to replication studies and specific to replication goals has not been clearly delineated. We believe that one contributor to some replication failures, and discrepancies across results more generally, is a lack of statistical power (or accuracy) for the desired goals. This situation is possible even when sample size is planned ahead of time, if the sample size planning approach is not in line to the particular goal or analysis. Though certainly not applicable to all replications, the detailed methods we discuss clearly show how researchers can believe that they are sufficiently powered in their attempt at replications but are, in fact, underpowered. Certainly, there are many other reasons for failures of replication, but the planning of replication studies has not caught up to the broadening appeal of replication goals in the literature and, correspondingly has seen too little use.

The methodological literature on sample size planning that can be used for the preceding replication goals is still in nascent stages, and more so for some domains than others. For example, if the design in question is an independent *t* test, there are many user-friendly software options available to conduct sample size planning. But some more complex designs are not yet implemented in software easy to use by nonspecialists or require researchers to hypothesize more complex relations and prespecify parameters. This is an active and growing area of research, but researchers can use general logic and steps outlined here to adapt sample size planning to other designs.

This article was not exhaustive in including all possible ways to measure and evaluate replication. Most generally, the focus was on direct replication studies. Although some of the strategies we described may be satisfactory for conceptual replications, given that there is always uncertainty in sample size planning, the additional variability from investigating a potentially different underlying population effect from the original study may encourage alternative methods. Other goals for direct replication are also found in the literature (see Table 1 for a variety of proposals). For example, despite the definitional and statistical power concerns raised earlier, researchers may want to test for a difference in effect size, given that sample values for replication effect sizes are often quite different from those reported in the original study. Alternatively, to address the most stringent definition of replication, that the magnitude of the effect size is equivalent in the original and replication studies, researchers can conduct an equivalence test for the difference in effect sizes (Anderson & Maxwell, 2016), though the required sample size would be quite large and uses of this goal might be limited in practice. Other thoughtful proposals have recently been advanced, particularly when multiple replication studies have been conducted, including statistically evaluating heterogeneity (Schauer & Hedges, 2021) and assessing the proportion of replications that exceed a minimally important effect size (Mathur & VanderWeele, 2019). Methodologists have recently began to advocate more strongly for multiple replication studies (e.g., Hedges & Schauer, 2019a), as we addressed earlier, and we agree with such proposals, which can harness the power of not only larger sample sizes, but also larger numbers of studies.

This article was divided into four different sections for sample size planning involving a single goal at a time. However, similarly to original research, researchers may have multiple goals for a replication study. First, a researcher may have multiple replication goals, such as, for example, demonstrating equivalence for a purported effect (ensuring that the CI falls completely within the equivalence region or rejecting both sides of a TOST) while also more accurately estimating its size (ensuring a sufficiently narrow CI width). Second, researchers may want to satisfy a particular goal in terms of power *and* accuracy (Kelley & Maxwell, 2003). Although Jiroutek et al. (2003) formulate an approach that jointly considers power and accuracy, similar composite methods are not yet developed for broader combinations of goals. We encourage researchers who have multiple goals to conduct multiple types of sample size planning and choose the (larger) sample size that will achieve both goals, or, at a minimum, learn that their selected sample size may only be optimal for particular goals. Generally, in many practical but certainly not always, the sample sizes necessary for Goals 2 and 3 will be larger than those needed for Goals 1 and 4, although there are a lot of caveats to this, including the true effect size, size of the equivalence region, desired CI width, and the statistical power and effect size of the relevant original study effect.

Finally, researchers should ideally consider replication as a process rather than a single event, where multiple studies are treated as pieces of a more cohesive scientific puzzle (e.g., Braver et al., 2014). Of course, once numerous replication studies have been conducted on a topic, the additional flexibility provided by meta-analytic methods and emphasized by Hedges and Schauer (2019a) becomes applicable, where the analyst is not collecting new data but aiming to have high statistical power and accuracy in both the number of studies and the sample sizes of the included studies. This article helps with the latter of these two goals, by encouraging researchers to collect replication study sample sizes that match the goal. We conclude by summarizing several of our main guidelines for success, organized by replication goal, in Table 3. We believe

Table 3

Summary of Guidelines for Replication Goals and Sample Size Planning

Goal	Guideline
1	• *Use when theory highlights existence and direction rather than magnitude*
	• Understand that nonsignificance does not directly imply original finding was a false positive
	• Do not use same sample size as original study by default
	• If basing effect size on reported original study effect, adjust for sampling uncertainty and publication bias using BUCSS (Anderson & Kelley, 2020) and/or options that address forms of uncertainty (e.g., McShane & Bockenholt, 2016; Pek & Park, 2019; Perugini et al., 2014)
	• If adjusted noncentrality parameter is zero using BUCSS, consider adjusting desired assurance or not using the original study for sample size calculations
	• If ignoring the original study, carefully choose theoretically relevant, minimally interesting effect size, and McShane and Bockenholt's (2014) method to adjust for effect size heterogeneity
2	• *Use when the goal is to demonstrate that the purported effect is so small as to be deemed trivial or null*
	• Can be used regardless of the statistical significance or effect size reported in the original study
	• Choose an equivalence bound small or narrow enough such that readers would agree such an effect is near-zero
	• If true parameter is thought to closely approximate zero, Chow et al.'s (2008) sample size formulas can be used for various two-group designs
	• If true parameter is thought to be small but nonzero, allocations or variances are unequal, or cost constraints are relevant, consider Guo et al.'s (2019) online sample size calculator for two-group designs
	• Understand that a large sample may be needed to demonstrate equivalence, in keeping with Brandolini's Law
3	• *Use when the goal is to accurately estimate the effect size, often improving upon poor estimation accuracy in the original study*
	• To plan sample size, use accuracy in parameter estimation (AIPE) procedures, implemented in MBESS
	• Carefully decide upon desired interval width, such that the range of plausible values is sufficiently narrow
	• Consider incorporating assurance for a higher degree of certainty that interval will be narrower than desired
	• When a fixed sample size is not needed in advance, particularly for parameters where the interval width depends on parameter size or when parameter estimate is thought to be highly uncertain, consider sequential AIPE, using the SMSD R package
4	• *Use when goal is to combine new knowledge (replication study) with prior results (original study, with possibly one or more previous replication studies), enhancing power and precision*
	• Carefully consider the assumptions for fixed-effect, random-effect, and varying-coefficient models
	• Conduct sample size planning for replication study based on desired accuracy of the combined effect size estimate
	• Use meta-analytic effect size estimate in sample size planning, adjusting for both estimation bias and publication bias, when possible
All	• Justify the selected replication goal, link the analysis to this goal, and provide details on corresponding sample size planning in the manuscript
	• When multiple goals are of interest, plan sample size for focal goal (with the knowledge that it may not be sufficient for all goals) or plan sample size for all goals and choose the largest feasible sample size
	• Consider the benefits of meta-analytically aggregating multiple replication studies
	• Particularly for more complex designs, keep apprised of future methodological articles, as this is an ongoing area of research and development

that this article is helpful for researches because it shows how to apply sophisticated replication sample size planning methods in easy-to-implement ways and distills the disparate literature into a unified treatment.

References

American Statistical Association Task Force. (2021). ASA President's task force statement on statistical significance and replicability. *Amstat News*. https://magazine.amstat.org/blog/2021/08/01/task-force-statement-p-value/

Anderson, S. F. (2020). Misinterpreting p: The discrepancy between p values and the probability the null hypothesis is true, the influence of multiple testing, and implications for the replication crisis. *Psychological Methods*, *25*(5), 596–609. https://doi.org/10.1037/met0000248

Anderson, S. F. (2021). Using prior information to plan appropriately powered regression studies: A tutorial using BUCSS. *Psychological Methods*, *26*(5), 513–526. https://doi.org/10.1037/met0000366

Anderson, S. F., & Kelley, K. (2020). *Bias-uncertainty corrected sample size* (1,2,0) [R Package]. https://cran.rproject.org/web/packages/BUCSS/index.htmlhttps://cran.rproject.org/web/packages/BUCSS/index.html

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. https://doi.org/10.1177/0956797617723724

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. https://doi.org/10.1037/met0000051

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

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*(2), 230–244. https://doi.org/10.1037/0022-3514.71.2.230

Beaton, D. E., Boers, M., & Wells, G. A. (2002). Many faces of the minimal clinically important difference (MCID): A literature review and directions for future research. *Current Opinion in Rheumatology*, *14*(2), 109–114. https://doi.org/10.1097/00002281-200203000-00006

Bonett, D. G. (2021). Design and analysis of replication studies. *Organizational Research Methods*, *24*(3), 513–529. https://doi.org/10.1177/1094428120911088

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods*, *13*(3), 173–181. https://doi.org/10.1037/a0012868

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, *14*(3), 225–238. https://doi.org/10.1037/a0016619

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, *15*(4), 368–385. https://doi.org/10.1037/a0020142

Bonett, D. G., & Price, R. M. (2015). Varying coefficient meta-analysis methods for odds ratios and risk ratios. *Psychological Methods*, *20*(3), 394–406. https://doi.org/10.1037/met0000032

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Publication bias. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis* (pp. 277–292). Wiley, Ltd. https://doi.org/10.1002/9780470743386.ch30

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Brown, A. W., Mehta, T. S., & Allison, D. B. (2017). Publication bias in science: What is it, Why is it Problematic, and How can it be Addressed? In K. H. Jamieson, D. Kahan, & D. A. Scheufele (Eds.), *The Oxford handbook of the science of science communication* (pp. 93–101). Oxford University Press.

Carter, E. C., & McCullough, M. E. (2018). A simple, principled approach to combining evidence from meta-analysis and high-quality replications. *Advances in Methods and Practices in Psychological Science*, *1*(2), 174–185. https://doi.org/10.1177/2515245918756858

Chattopadhyay, B., & Kelley, K. (2017). Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation. *Psychological Methods*, *22*(1), 94–113. https://doi.org/10.1037/met0000089

Chow, S.-C., Shao, J., & Wang, H. (2008). *Sample size calculations in clinical research* (2nd ed., Vol. 11). CRC Press. https://doi.org/10.1201/9780203911341

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*(3), 243–253. https://doi.org/10.1037/1082-989X.8.3.243

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Dallow, N., & Fina, P. (2011). The perils with the misuse of predictive power. *Pharmaceutical Statistics*, *10*(4), 311–317. https://doi.org/10.1002/pst.467

Dawes, R. M. (2004). Commentary on Meehl's theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Applied & Preventive Psychology*, *11*(1), 23–25. https://doi.org/10.1016/j.appsy.2004.02.002

DeMets, D. L., & Lan, K. K. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, *13*(13-14), 1341–1352. https://doi.org/10.1002/sim.4780131308

Dupont, W. D., & Plummer, W. D., Jr. (1990). Power and sample size calculations. A review and computer program. *Controlled Clinical Trials*, *11*(2), 116–128. https://doi.org/10.1016/0197-2456(90)90005-M

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. https://doi.org/10.1016/j.jesp.2015.07.009

FDA. (2001). *Guidance for industry on statistical approaches to establishing bioequivalence*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/statistical-approaches-establishing-bioequivalence

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Psychology*, *58*(2), 203–210. https://doi.org/10.1037/h0041593

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Sansone, C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review*, *18*(1), 3–12. https://doi.org/10.1177/1088868313507536

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. https://doi.org/10.1177/1745691614551642

Gelman, A., & Loken, E. (2014). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *American Scientist*, *102*, 460–465. https://doi.org/10.1511/2014.111.460

Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician*, *60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science”. *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Gliklich, R. E., Dreyer, N. A., & Leavy, M. B. (2014). *Analysis, interpretation, and reporting of registry data to evaluate outcomes. in registries for evaluating patient outcomes: A user’s guide* (3rd ed.). Agency for Healthcare Research and Quality. https://www.ncbi.nlm.nih.gov/books/NBK208602/

Goodman, S. N. (2016). STATISTICS. Aligning statistical and scientific reasoning. *Science*, *352*(6290), 1180–1181. https://doi.org/10.1126/science.aaf5406

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

Guo, J. H., Chen, H. J., & Luh, W. M. (2019). Optimal sample sizes for testing the equivalence of two means. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *15*(3), 128–136. https://doi.org/10.1027/1614-2241/a000171

Haaf, J. M., Ly, A., & Wagenmakers, E.-J. (2019). Retire significance, but still test hypotheses. *Nature*, *567*(7749), 461–461. https://doi.org/10.1038/d41586-019-00972-7

Hahn, G. J., & Meeker, W. Q. (1991). *Statistical intervals: A guide for practitioners* (1st ed.). Wiley-Interscience. https://doi.org/10.1002/9780470316771

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*(2), 490–499. https://doi.org/10.1037/0033-2909.92.2.490

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85. https://doi.org/10.2307/1164832

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88*(2), 359–369. https://doi.org/10.1037/0033-2909.88.2.359

Hedges, L. V., & Schauer, J. M. (2019a). More than one replication study is needed for unambiguous tests of replication. *Journal of Educational and Behavioral Statistics*, *44*(5), 543–570. https://doi.org/10.3102/1076998619852953

Hedges, L. V., & Schauer, J. M. (2019b). Statistical analyses for studying replication: Meta-analytic perspectives. *Psychological Methods*, *24*(5), 557–570. https://doi.org/10.1037/met0000189

Held, L., Pawel, S., & Schwab, S. (2020). Replication power and regression to the mean. *Significance*, *17*(6), 10–11. https://doi.org/10.1111/1740-9713.01462

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*(3), 315–332. https://doi.org/10.1037/1082-989X.5.3.315

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). SAGE Publications, Inc. https://doi.org/10.4135/9781412985031

Jiroutek, M. R., Muller, K. E., Kupper, L. L., & Stewart, P. W. (2003). A new method for choosing sample size for confidence interval-based inferences. *Biometrics*, *59*(3), 580–590. https://doi.org/10.1111/1541-0420.00068

Kafadar, K. (2021). EDITORIAL: Statistical significance, P-values, and replicability. *The Annals of Applied Statistics*. Advance online publication. https://doi.org/10.1214/21-AOAS1500

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24. https://doi.org/10.18637/jss.v020.i08

Kelley, K. (2020). *Sequential Methods for Study Design* [R Package]. https://github.com/yelleKneK/SMSD

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. https://doi.org/10.1037/1082-989X.8.3.305

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*(4), 363–385. https://doi.org/10.1037/1082-989X.11.4.363

Kelley, K., Bilson Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kelley, K., Bilson Darku, F., & Chattopadhyay, B. (2019). Sequential accuracy in parameter estimation for population correlation coefficients. *Psychological Methods*, *24*(4), 492–515. https://doi.org/10.1037/met0000203

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kieser, M., & Hauschke, D. (1999). Approximate sample sizes for testing hypotheses about the ratio and difference of two means. *Journal of Biopharmaceutical Statistics*, *9*(4), 641–650. https://doi.org/10.1081/BIP-100101200

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kuffner, T. A., & Walker, S. G. (2019). Why are p-values controversial? *The American Statistician*, *73*(1), 1–3. https://doi.org/10.1080/00031305.2016.1277161

Lai, K., & Kelley, K. (2011). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods*, *16*(2), 127–148. https://doi.org/10.1037/a0021764

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

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical & Statistical Psychology*, *31*(2), 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Lazarou, I., Parastatidis, T., Tsolaki, A., Gkioka, M., Karakostas, A., Douka, S., & Tsolaki, M. (2017). International ballroom dancing against neurodegeneration: A randomized controlled trial in Greek community-dwelling elders with mild cognitive impairment. *American Journal of Alzheimer's Disease and Other Dementias*, *32*(8), 489–499. https://doi.org/10.1177/1533317517725813

Lehmann, E. L., & Romano, J. P. (2005). *Testing statistical hypotheses* (3rd ed.). Springer-Verlag. https://doi.org/10.1007/0-387-27605-X

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Lindsay, R. M., & Ehrenberg, A. S. C. (1993). The design of replicated studies. *The American Statistician*, *47*(3), 217–228. https://doi.org/10.2307/2684982

Lipsey, M. W., & Aiken, L. S. (1990). *Design sensitivity: Statistical power for experimental research*. Sage.

Liu, X. S. (2015). Sample size and the precision of the confidence interval in meta-analyses. *Therapeutic Innovation & Regulatory Science*, *49*(4), 593–598. https://doi.org/10.1177/2168479015570332

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology*, *20*(1), 76–78. https://doi.org/10.1037/0278-6133.20.1.76

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Manapat, P. D., Anderson, S. F., & Edwards, M. C. (2022). A revised and expanded taxonomy for understanding heterogeneity in research and reporting practices. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000488

Mathur, M. B., & VanderWeele, T. J. (2019). Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges and Schauer (2019). *Psychological Methods*, *24*(5), 571–575. https://doi.org/10.1037/met0000223

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). Routledge.

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*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

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. https://doi.org/10.1037/a0039400

Mayo, D. G. (2020). P-values on trial: Selective reporting of (best practice guides against) selective reporting. *Harvard Data Science Review*. Advance online publication. https://doi.org/10.1162/99608f92.e2473f6a

Mayo, D. G. (2021). Significance tests: Vitiated or vindicated by the replication crisis in psychology? *Review of Philosophy and Psychology*, *12*(1), 101–120. https://doi.org/10.1007/s13164-020-00501-w

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling*, *23*(5), 750–773. https://doi.org/10.1080/10705511.2016.1186549

McShane, B. B., & Bockenholt, U. (2014). You cannot step into the same river twice: When power analyses are optimistic. *Perspectives on Psychological Science*, *9*(6), 612–625. https://doi.org/10.1177/1745691614548513

McShane, B. B., & Bockenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods*, *21*(1), 47–60. https://doi.org/10.1037/met0000036

McShane, B. B., Bockenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730–749. https://doi.org/10.1177/1745691616662243

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

McShane, B. B., Tackett, J. L., Böckenholt, U., & Gelman, A. (2019). Large-scale replication projects in contemporary psychological research. *The American Statistician*, *73*(Suppl. 1), 99–105. https://doi.org/10.1080/00031305.2018.1505655

Moher, J. (2020). Distracting objects induce early quitting in visual search. *Psychological Science*, *31*(1), 31–42. https://doi.org/10.1177/0956797619886809

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*(6), 1289–1290. https://doi.org/10.1177/0956797614525969

NCSS LLC. (2021). *PASS 2021 Power analysis and sample size software*. ncss.com/software/pass

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLoS ONE*, *11*(6), e0157732. https://doi.org/10.1371/journal.pone.0157732

Nordgren, L. F., & McDonnell, M.-H. M. (2011). The scope-severity paradox: Why doing more harm is judged to be less harmful. *Social Psychological & Personality Science*, *2*(1), 97–102. https://doi.org/10.1177/1948550610382308

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*(3), 187–201. https://doi.org/10.1002/pst.175

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Eds.' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. https://doi.org/10.1177/1548051815614321

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should we expect when we replicate? A statistical view of replicability in psychological science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Pek, J., & Park, J. (2019). Complexities in power analysis: Quantifying uncertainties with a Bayesian-classical hybrid approach. *Psychological Methods*, *24*(5), 590–605. https://doi.org/10.1037/met0000208

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on*

Psychological Science: A Journal of the Association for Psychological Science, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

Pornprasertmanit, S., & Schneider, W. J. (2014). Accuracy in parameter estimation in cluster randomized designs. *Psychological Methods*, *19*(3), 356–379. https://doi.org/10.1037/a0037036

Rao, T. S. S., & Andrade, C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian Journal of Psychiatry*, *53*(2), 95–96. https://doi.org/10.4103/0019-5545.82529

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, *146*(8), 701–719. https://doi.org/10.1037/bul0000232

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, *55*(3), 182–186. https://doi.org/10.1198/000313001317097960

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe & Huber.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, *40*(1), 73–83. https://doi.org/10.1037/0003-066X.40.1.73

Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PLoS ONE*, *11*(9), e0162093. https://doi.org/10.1371/journal.pone.0162093

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

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. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics. Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

Uy, J. (2019, March 14). Brandolini's Law, Emotional certainties, and effective bullshit detection. *Medium*. https://medium.com/big-on-development/brandolinis-law-emotional-certainties-and-effective-bullshit-detection-4a605eb4a4db

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 215–247. https://doi.org/10.3102/1076998609346961

van Ravenzwaaij, D., & Wagenmakers, E.-J. (2021). Advantages masquerading as "issues" in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000415

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. https://doi.org/10.1037/a0036731

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggin, B., & Grahe, J. E. (2019). Publishing research With undergraduate students via replication work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*, 247. https://doi.org/10.3389/fpsyg.2019.00247

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < .05$". *The American Statistician*, *73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, *15*(1), 18–37. https://doi.org/10.1037/a0015917

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Zhang, P. (2003). A simple formula for sample size calculation in equivalence studies. *Journal of Biopharmaceutical Statistics*, *13*(3), 529–538. https://doi.org/10.1081/BIP-120022772

Received August 23, 2021
Revision received May 5, 2022
Accepted May 23, 2022 ■