Planning Falsifiable Confirmatory Research

James E. Kennedy

Accepted for publication in *Psychological Methods*.

Author Note

I thank Professor Caroline Watt for reviewing an earlier version of this paper and making helpful comments.

I have no known conflict of interest to disclose. Earlier drafts of this paper were posted on PsyArXiv (https://doi.org/10.31234/osf.io/pu2xy) and on the author's personal website.

Correspondence concerning this article should be addressed to James E. Kennedy, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 7 George Square, Edinburgh, Scotland EH8 9JZ.
Email: jek@jeksite.org

Abstract

Falsifiable research is a basic goal of science and is needed for science to be self-correcting. However, the methods for conducting falsifiable research are not widely known among psychological researchers. Describing the effect sizes that can be confidently investigated in confirmatory research is as important as describing the subject population. Power curves or operating characteristics provide this information and are needed for both frequentist and Bayesian analyses. These evaluations of inferential error rates indicate the performance (validity and reliability) of the planned statistical analysis. For meaningful, falsifiable research, the study plan should specify a minimum effect size that is the goal of the study. If any tiny effect, no matter how small, is considered meaningful evidence, the research is not falsifiable and often has negligible predictive value. Power ≥.95 for the minimum effect is optimal for confirmatory research and .90 is good. From a frequentist perspective, the statistical model for the alternative hypothesis in the power analysis can be used to obtain a *p* value that can reject the alternative hypothesis, analogous to rejecting the null hypothesis. However, confidence intervals generally provide more intuitive and more informative inferences than *p* values. The preregistration for falsifiable confirmatory research should include (a) criteria for evidence the alternative hypothesis is true, (b) criteria for evidence the alternative hypothesis is false, and (c) criteria for outcomes that will be inconclusive. Not all confirmatory studies are or need to be falsifiable.

Keywords: falsification, power analysis, confirmatory research, falsifiable hypothesis, operating characteristics

Translational Abstract

Science is based on developing empirically testable theories or hypotheses. This implies that research can provide evidence about whether a scientific theory or hypothesis is true or false. However, past research practices in psychology have focused on finding evidence that a hypothesis is true with little attention to the methods for evidence that the hypothesis may be false. The statistical methods for conducting falsifiable research are not well known in the social sciences. The purpose of this paper is to bring together various concepts and methods to form a useful framework for conducting research that can provide evidence that a scientific hypothesis is false as well as true. For falsifiable research, scientists must pre-specify a study outcome or magnitude of effect that would be considered evidence that the hypothesis is not true. If any obtained effect, no matter how close to zero, could be considered evidence supporting the hypothesis, the hypothesis is not falsifiable—and therefore is of questionable scientific value. Power analysis or operating characteristics are used to design studies that can provide evidence that a hypothesis is true or false. These study design practices are needed for Bayesian analyses as well as for traditional frequentist statistical analysis. Preregistered confirmatory research with a pre-specified minimum effect size for study design and appropriate power analysis are required for falsifiable scientific theories or hypotheses. Studies with more than 1000 subjects will often be needed.

Planning Falsifiable Confirmatory Research

Research methods that can provide evidence that a scientific hypothesis is true but cannot provide evidence that the hypothesis is false are biased and undermine the assumption that science is self-correcting. Science is based on testable hypotheses, which implies a useful degree of falsifiability.

The "replication crisis" or "methodological revolution" in psychological research (Nelson et al., 2018; Wagenmakers, 2015) resulted from a combination of research practices that could provide positive but not negative evidence about the researchers' hypotheses. These practices were common even though textbooks about statistics described falsifiable research as a requirement for science and as distinguishing science from pseudoscience (Goodwin, 2010; Jackson, 2011).

The failure to distinguish between exploratory research and confirmatory research was a major underlying cause of the replication crisis (Wagenmakers et al., 2012). Exploratory research typically has and should have much flexibility for researchers to adapt the analyses and/or hypotheses during data analysis (Hoaglin et al., 1983). Such practices are expected and appropriate, but tend to produce biased findings that must be verified with confirmatory research that does not have exploratory flexibility (Gelman & Loken, 2014; Hoaglin et al., 1983, pp. 1-2; Kerr, 1998; Simmons et al., 2011; Wagenmakers et al., 2012). Exploratory research is an important, creative part of the scientific process. However, presenting exploratory findings as if they were preplanned seriously distorts scientific evidence.

Exploratory research typically cannot and is not intended to provide evidence that a hypothesis is false. The focus is on finding an effect. The most common strategy is that data from a relatively small sample size is searched for relatively large effects. The absence of an effect could be due to small sample size rather than because the hypothesis is false. In addition, measurement methods and experimental manipulations are often being developed and tested in exploratory research. Failure to find an effect may be due to inadequate measurement methods or inadequate experimental manipulations, rather than because the effect is false. Exploratory research that does not find evidence for an effect has often not been reported, which creates a positive bias in the literature. Prior to 2012, most research in psychology appears to have been

conducted with exploratory methods (Nelson et al., 2018; Wagenmakers et al., 2012). Formal confirmatory research was rare.

Statistical analyses that are intrinsically unfalsifiable have been another factor inhibiting falsifiable research. The typical null hypothesis tests used in psychological research in past decades have predicted only that the magnitude of the effect "is not zero" or will be "not due to chance." These hypotheses do not predict the size of the effect and are not falsifiable in principle because any finite sample size could have inadequate power to detect the tiny effects consistent with the hypotheses.

The traditional recommendation for statistical power of .80 is also not adequate for evidence that a hypothesis is false. A power of .80 has .20 probability of incorrectly obtaining a nonsignificant result when the effect is true. Here too, a significant outcome is interpreted as evidence for the effect, but a nonsignificant outcome is inconclusive rather than evidence that the effect is false. Although many confirmatory studies in recent years have appropriately had a power of .90 or higher, a design power of .80 is still common based on guidelines from past decades before formal confirmatory research was conducted in psychology.

The replication crisis has stimulated many discussions and debates about appropriate statistical methods (Mayo, 2018). However, few of these discussions have distinguished between exploratory and confirmatory research. For example, in 2019 the *American Statistician* journal devoted a special issue to the topic of statistical inference, that included a summary editorial (Wasserstein et al., 2019) and 43 articles (401 pages). Most of the authors did not distinguish between exploratory and confirmatory research. The articles primarily discussed proposed statistical methods to address the biases in exploratory research without noting that the biases will not occur in properly designed confirmatory research.

Tong (2019) argued in the special issue that the statistical problems will not be resolved until the distinction between exploration and confirmation becomes a fundamental component of statistical thinking. He pointed out that *p* values and hypothesis tests are applicable for preregistered confirmatory research, but not for exploratory research. The same point was noted in the classic book about exploratory analyses by Hoaglin, Mosteller, & Tukey (1983, pp. 1-2).

Discussions and guidelines about study preregistration emphasize the distinction between exploration and confirmation (Center for Open Science, n.d.; Wagenmakers et al., 2012;

Wicherts et al., 2016). Unfortunately, most discussions about statistical methods appear to have not kept up with the current research practices.

Properties of Confirmatory Research

Confirmatory research focuses on testing a prespecified prediction. Confirmatory predictions are not applicable for all research. For example, public opinion surveys and studies of disease prevalence seek descriptive information and can be conducted without a prediction, and without the expectation of confirmation.

The terms confirmatory and exploratory are used differently by different researchers. Some researchers apply the term confirmatory only to confirmation of a previous empirical finding. For these researchers, the initial research testing a theoretical hypothesis is described as exploratory. However, other researchers apply the term confirmatory to the initial research testing (confirming) a theoretical hypothesis. For these researchers, exploratory research occurs when researchers are searching for and do not yet have a specific hypothesis in mind. Thus, one researcher may classify a study as confirmatory that another would classify as exploratory.

The position in this paper is that the distinction between exploration and confirmation is based on certain methodological properties that are applicable whether or not previous research has been conducted. Good confirmatory research will have these properties.

The fundamental and minimum property for confirmatory research is that a prespecified prediction is investigated using study plans that are prespecified in sufficient detail to eliminate researcher flexibility to adapt the analyses and/or hypotheses after looking at the data. The research should also have measurement methods and experimental manipulations that are reasonably well established. If the measurement methods or experimental manipulations are being developed as part of the study, the study is more appropriately considered exploratory.

Falsifiable confirmatory research will also have a planned statistical analysis and appropriate sample size that can provide evidence that a hypothesis is false as well as true. Not all confirmatory studies are or need to be falsifiable. *Underpowered confirmatory research* that does not have an adequate sample size for falsifiable research may be valuable, particularly if it can be included in a meta-analysis of preregistered studies that eliminated researcher flexibility for the effect of interest.

Most confirmatory studies will also include preplanned and post hoc exploratory analyses. The exploratory analyses in a study powered for falsifiable confirmatory research will usually have large sample sizes.

Exploratory research can have many different purposes and associated analysis methods. These purposes include searching for an effect or hypothesis, developing measurement methods and experimental manipulations, developing methods for analyzing the data, verifying that participants can be recruited, verifying that instructions and data collection software are effective, and testing a hypothesis using fully prespecified methods that would be appropriate for confirmatory research, but the researchers consider the hypothesis tentative or exploratory. The latter case may be described as *fully prespecified exploratory research*, and may be appropriate to include in a meta-analysis of preregistered research that excluded researcher flexibility.

Several exploratory studies may be useful in developing and debugging the methodology for a larger confirmatory study. Unanticipated problems during the conduct and analyses of an experiment can be symptoms of premature confirmatory research.

Two-sided analyses are based on an ambiguous prediction (either a positive or a negative effect) and are typically used at the exploratory stage of searching for an effect or hypothesis. Unambiguous predictions are preferred for confirmatory research. However, in certain situations, a two-sided confirmatory analysis may be applicable. For example, different researchers could have conflicting hypotheses or conflicting previous findings for an experimental manipulation.

Falsifiable Research Methods

Although falsifiable research is a fundamental goal of science, the statistical methods for conducting falsifiable research are not well known in the social sciences. Relevant writings are scattered and diverse, with key writings outside the psychological literature. Also, methodological papers usually discuss one topic in isolation, such as preregistration or noncentral statistical distributions, without integrating or discussing the relationships among the various topics that are needed to conduct research. This piecemeal approach hinders understanding and implementation of research methodology.

The purpose of this paper is to integrate relevant topics into a framework for understanding and implementing research that can provide evidence that a hypothesis is false as well as true. This paper is not intended to be a comprehensive review of the ongoing debates

about statistical methodology or the history of the methods described here. The goal is more practical: to provide a useful framework and methods for researchers who want to conduct falsifiable confirmatory research. The paper is intended to be useful to a wide variety of readers, including those who consider statistical methods as a necessary tool for research, but not a primary interest.

This paper focuses on simple study designs that would be analyzed with simple statistics such as t-tests. This allows basic principles to be described with minimal distractions. In addition, such simple methods are frequently used in practice, and are particularly appropriate for confirmatory research with well-developed predictions. Some of the specific methods discussed here may need to be adapted for more complex research, but the basic principles will still apply.

Practical falsifiable confirmatory research methods involve five steps.

1. The primary hypothesis of interest, associated research methods, and analysis plan are preregistered as confirmatory and are not subject to exploratory flexibility. Exploratory and post hoc analyses will usually be done, but are clearly distinguished from the confirmatory analyses. The measurement instruments and experimental manipulations have reasonably established validity and reliability from previous (often exploratory) research.
2. At the planning or *predata* stage, the inferential error rates for the planned confirmatory analysis are evaluated for different possible true effect sizes to determine what effect sizes can be confidently investigated in the study. These error rates or power curve evaluate the performance (validity and reliability) of the planned statistical analysis for different effect sizes and are needed for both frequentist and Bayesian statistics. The optimal practice for falsifiable research is to specify a minimum effect size for study design, and plan the study to investigate that effect size with high power. A sometimes necessary alternative is to set the sample size based on limited available resources.
3. The preregistered analysis plan includes the specific statistical tests and the criteria that will be considered: (a) evidence that the alternative hypothesis is true, (b) evidence that the alternative hypothesis is false, and (c) inconclusive. Confidence intervals and other

analyses that will be used as part of the inferences are also specified in the preregistration with corresponding criteria.

4. The confirmatory outcome of the study is evaluated in accordance with the preregistered analysis plan.
5. Prior to or during review for publication, the preregistration is compared to the draft report of the study to verify that the preregistration was sufficiently detailed to prevent researcher flexibility, that the study was conducted in accordance with the preregistration, and that any deviations from the preregistration were identified and explained. A few such protocol deviations can be expected for a large study.

The Transparent Psi Project is an example of research that implemented all five steps above (Kekecs et al., 2023).

The present paper addresses steps 2, 3 and 4. These steps apply for both frequentist and Bayesian statistics. First, a discussion of the goals for falsifiable research.

Realistic Goals for Falsifiable Statistical Research

The arguments that replicable findings cannot be expected in social psychology appear to undermine the concept of falsifiable research. This section discusses those arguments, including reasonable expectations for falsifiable research.

A statistical hypothesis test is based on collecting data to provide evidence about a mathematical statistical model or statistical hypothesis that in turn provides evidence about a theoretical hypothesis. The data provide evidence about the theoretical hypothesis to the extent that the statistical model is valid, reliable, and accurately represents or implements the theoretical hypothesis.

Reported evidence that the statistical model is true or false is not necessarily corresponding evidence about the theoretical hypothesis. In addition to the replication-crisis methodological problems of undisclosed biases, selections, and low power (unreliability) during data collection and analysis, the measures and/or experimental manipulations may not usefully represent the theoretical hypothesis. This is a significant concern in psychological research (Earp & Trafimow, 2015; Meehl, 1990; Mayo, 2018, pp. 92-106).

In practice, the question that is answered true or false in confirmatory scientific research is “do scientists adequately understand the theoretical hypothesis and associated research methods to reliably make accurate verifiable predictions?” An inability to demonstrate reliable predictions may be due to inadequate understanding of appropriate research methods rather than because the theoretical hypothesis is false. Scientific evidence requires both.

Psychological research is particularly susceptible to doubts about the research methods. Any differences in subject population, background culture, measurement methods, study procedure, or study environment could influence the occurrence and magnitude of an effect. Given that a replication study will inevitably have differences from previous studies, some authors have argued that high rates of successful replications cannot be expected in psychological research, particularly in social psychology (Amrhein et al., 2019; Cesario, 2014; Gergen, 1973).

Gergen (1973, 1994) took these points to the logical conclusion that social psychology should be classified as a branch of history rather than science. He argued that human interactions are based on continually changing personal dispositions and culture. Replicable effects and useful predictions cannot be expected, just reports of events as or after they occurred. According to Gergen (1973), the scientific method may be appropriate when “certain phenomena may be closely tied to physiological givens” (p. 318).

Confirmatory research is based on the premise that stable, predictable effects can be discovered. If researchers believe that reliably replicable effects or predictions are essentially impossible in a certain area of research, Gergen’s arguments about history versus science would appear to be applicable.

In context of falsifiable research as discussed in this paper, the phrase “evidence that the hypothesis is false” means that the outcome of the statistical analysis reached the criterion that was pre-specified for evidence that the statistical model of the alternative hypothesis is not true for the conditions of the study. This is evidence that the researchers do not correctly understand the theoretical hypothesis and/or the associated research methods.

Providing evidence that a hypothesis is true or false does not imply that one study can provide definitive evidence or fully resolve whether a hypothesis is true or false. Confident inferences must be based on multiple confirmatory studies by different researchers (Earp &

Trafimow, 2015; Mayo, 2018) and converging evidence. Each confirmatory study contributes to the overall evidence for a line of research and should to the extent possible provide strong unbiased evidence.

The rest of this paper focuses on the statistical model for a predicted effect, and puts aside the important question of whether the statistical model appropriately represents the researchers' theoretical hypothesis. The terms falsifiable prediction and falsifiable hypothesis are used interchangeably in this paper to mean that a study can provide evidence that a statistical hypothesis is false as well as true.

Minimum effect size for study design

Describing the effect sizes that can be reliably investigated in a planned confirmatory study is basic study design information, similar to describing the subject population. As noted in the introduction, if researchers are not willing to specify an effect size that will be evidence that the hypothesis is false or negligible, falsifiable research is not possible. In that case, research studies can only provide evidence that supports the hypothesis or is inconclusive. This situation compromises the basic principles of science.

The term *minimum effect size for study design* is used here for the smallest effect size that the researchers explicitly or implicitly hope to obtain in a study. This effect size can be based on effects found in previous research, on a smallest effect size of practical or theoretical interest, or on a sample size that can be obtained with the available resources that the investigators are willing to commit to the project. Researchers planning a study with sample size based on available resources may not explicitly specify a minimum effect size for study design. However, effect sizes that are too small to be reliably detected in the study are below an implied minimum effect size for study design. The effect sizes that can and cannot be confidently investigated with a planned study can be determined by a power curve or operating characteristics as discussed in the next section. Minimum effect sizes for study design apply for Bayesian hypothesis tests as well as for frequentist tests.

For falsifiable confirmatory research, evidence that the true effect is smaller than the minimum effect size for study design is interpreted as evidence that the effect is negligible or false. Recent writings often use the term smallest effect size of interest (SESOI) in contexts that

can mean either a minimum effect size for study design or a smallest effect size of practical or theoretical interest. For the purposes of this paper, I find it useful to distinguish these cases and have not used the SESOI terminology here.

As noted above, a confirmatory hypothesis test will usually be one-sided. For general linear models or ANOVAs, more powerful directional contrasts for confirmatory hypotheses are preferred rather than the intrinsically two-sided and more exploratory ANOVA main effects and interactions. Rosenthal, Rosnow, & Rubin (2000) explain the rationale, benefits, and methods for implementing contrasts.

Effect Sizes Found in Previous Research

Planning a confirmatory study based on effect sizes or sample sizes in previous research is fraught with challenges and pitfalls. As described in the introduction, exploratory research typically has flexible methodology that produces exaggerated effect sizes. Any analysis that was not preregistered as confirmatory can reasonably be assumed to have flexible exploratory methodology. Available evidence indicates that retrospective meta-analyses do not adequately compensate for the exploratory biases and tend to produce inflated effect sizes (Kvarven et al., 2019; Nelson et al., 2018). Initial studies also tend to have small sample sizes and associated wide confidence intervals for effect sizes. In addition, large heterogeneity or between-study variation in effect sizes due to unknown factors is common in psychological research and tends to make power estimates overly optimistic, particularly for small effects (McShane & Böckenholt, 2014; Stanley et al., 2018). Virtually all writings about power analysis conclude that the mean effect size in previous research is not appropriate for planning a confirmatory study in psychology.

Perugini, Gallucci, and Costantini (2014) proposed a simple “safeguard” power analysis that is based on the lower 80% or 95% one-sided confidence interval for the effect size estimate from previous research. Other more complicated and sometimes less conservative strategies have been proposed (e.g., Anderson, Kelley & Maxwell, 2017; Anderson, & Maxwell, 2017; McShane & Böckenholt, 2016). However, none of these proposals have attempted to incorporate full adjustments for the inflated estimates from exploratory research methodology. These methods may be applicable when based on previous preregistered confirmatory research.

Useful guidelines for estimating the magnitude of effect size inflation in previous exploratory and unregistered research remain to be developed. Kvarven, Strømland, and Johannesson (2019) found that the average effect size from 15 retrospective meta-analyses was almost three times larger than the effect size in subsequent preregistered multi-lab confirmatory studies. Similarly, the average effect size from 100 published studies was about two times larger than the average effect size in the subsequent preregistered confirmatory studies (Open Science Collaboration, 2015). The median effect size in previous studies was four times larger for another 26 preregistered multi-lab confirmatory studies (Klein et al., 2018). Stanley, Doucouliagos, and Ioannidis (2022) compared preregistered well-powered confirmatory research with previous studies and meta-analyses. They concluded that a false-positive rate as high as 50% is consistent with available data. They also concluded that meta-analyses based on larger studies with "median retrospective power" above 50% are more likely to confirm successfully than meta-analyses based primarily on smaller studies. The latter "should be interpreted with great caution or discounted altogether" (page 88). For example, they reported that a meta-analysis of Bem's (2011) widely discussed studies of precognition had a median retrospective power of 7.7% (page 97).

Basing a planned study on the effect sizes found in previous research ultimately focuses on the question: is the previous research valid? However, an effect may be real even if the previous research is invalid or greatly exaggerated. The more relevant scientific question is: does a meaningful effect occur?

Smallest Effect Size of Practical or Theoretical Interest

The most frequently recommended strategy for planning confirmatory research is to identify a smallest effect size of practical or theoretical interest for the specific topic being investigated. A smaller effect would be considered negligible or not worth pursuing. This strategy is not affected by inflated effects in previous exploratory research. Kruschke's (2015) "Region of Practical Equivalence (ROPE)" for Bayesian analysis is the same concept.

Although this strategy has been discussed for about seven decades, the identification and implementation of a smallest effect size of practical or theoretical interest remains rare outside of applied research (Cohen, 1965, 1988; Hodges & Lehmann, 1954; Lakens et al., 2018; Lenth, 2001; Serlin & Lapsley, 1993). An underlying problem is that when a hypothesis test provides

evidence that the data are not consistent with the null model (whether by frequentist or Bayesian methods), an argument that the effect is not meaningful can be expected to be controversial. As Cohen (1965) noted:

> In much "pure" research, in principle, *any* nonzero effect, no matter how small, may be meaningful in relation to a theory: i.e., no negligible effects exist. This would limit the applicability of the idea … to research, mostly "applied," where negligible effects can be meaningfully defined. (page 101)

At the same time, it has long been recognized that trivial effects with no practical consequences can be statistically significant with large sample sizes (Greenland, 2019; Ioannidis, 2005; Pogrow, 2019; Stanley et al., 2022). In such cases, "… little or nothing actually is contributed to our ability to predict one thing from another" (Hays, 1994, page 335).

Opinions about the implications of small effects in psychological research vary widely. Some writers have identified situations in which small effects have significant implications and argued that attention to small effects should be common (Funder & Ozer, 2019; Götz et al., 2022; Rosnow & Rosenthal, 2003). Others have argued that most small effects do not have practical consequences and should be ignored (Pogrow, 2019; Primbs et al., 2023; Stanley et al., 2022). All writers agree that small effects can have value in some situations and not others. The debate appears to be about whether meaningful small effects should be presumed to occur frequently in psychological research, and whether the burden of proof resides with those who presume that a small effect is or is not meaningful.

A highly relevant observation is that "[t]he smaller the effect sizes in a scientific field, the less likely the research findings are to be true" (Ioannidis, 2005, p. 0697; also, Pogrow, 2019). This point is supported by both logic and by evidence in psychological research (Mitchell, 2012; Open Science Collaboration, 2015; Stanley et al., 2022). In addition to researcher flexibility, well known sources of bias that are a greater concern for smaller effect sizes include demand characteristics, experimenter expectancy effects, and common method bias (Klein et al., 2012; Podsakoff et al., 2003). In a review of expectation bias, Jussim (2017) concluded that exaggerated effect sizes have been frequently reported for the biases, but small biases do occur. Correlation coefficients in the range of r = .06-.20 were typical magnitudes for the small biases. Potential biases from the handling of incomplete or noncompliant data have been less frequently

discussed, but are a threat that should be addressed in virtually all research (for example, Bouwmeester et al., 2017). For correlational research with large databases, Ferguson and Heene (2021) suggested that examining the magnitudes of correlations with nonsensical variables (not meaningfully causal) can provide a context for small effects. For the database they investigated, they concluded that statistically significant correlations of less than $r = .10$ should be treated as noise.

Cohen (1965, 1988) tentatively suggested that an effect size equivalent to a correlation coefficient of .10 (accounting for 1% of the variance) or the equivalent Cohen's *d* of .20 might be a small effect that could be used for setting sample size in psychological research "when no better basis for estimating the [effect size] is available" (Cohen, 1988, p. 25). He made this suggestion with substantial reservations and made clear that a better rationale for small effects was much preferred when possible.

Commenting on Cohen's criterion, Stanley, Doucouliagos, and Ioannidis (2022, page 103, endnote xiii) observed that "[c]urrent methods, protocols, and measures of psychological experiments are generally not sufficiently precise to reliably detect an effect associated with less than 1% of the observed outcome variation." They noted that effects smaller than Cohen's criterion are usually associated with low median retrospective power and are an indication that a finding is likely false or at best scientifically trivial. The conclusions of Stanley, Doucouliagos, and Ioannidis (2022) are based on exploratory and unregistered research. Their applicability to preregistered confirmatory research remains to be evaluated when more studies are available.

Schäfer and Schwarz (2019) concluded that neither Cohen's criteria nor the common recommendation to consider other effect sizes in the area of research are useful at present. They pointed out that the striking differences in results between studies with and without preregistration indicate that a much larger database of preregistered studies is needed before any guidelines about effect size can be developed.

An initial attempt to develop expert opinion about a smallest effect size of interest in a subject area was not successful in developing a consensus (Riesthuis et al., 2021). The authors interpreted this as indicating that the development of a smallest effect size of practical or theoretical interest may be more useful for an individual effect than for a general area of

research. Anvari and Lakens (2021) proposed an empirical approach that may be useful when the smallest effect size of interest can be based on the smallest subjectively experienced difference.

Recommendations about Smallest Effect Size of Practical or Theoretical Interest

Greater skepticism about smaller effects appears to be justified given the currently available evidence that small effects are less likely to be true as well as less likely to be meaningful. The possibility of methodological bias should receive careful attention for very small effects. Arguments that a small effect is meaningful should be based on the specific hypothesis being investigated, not just the fact that very small effects are considered meaningful in certain other areas of science. Such arguments about other areas of science do not provide evidence about whether the particular small effect is meaningful and overlook the fact that very small effects are not considered meaningful in some areas of science.

For confirmatory research that is intended to have implications for applied use, it seems reasonable to expect that the effect sizes that will be considered to have practical significance for the application of the research can be described at the study planning stage and incorporated into the power analysis and preregistration.

The problem case is initial confirmatory research that is theoretically oriented rather than applied, in situations without previous confirmatory (or fully prespecified exploratory) studies and without an apparent rationale for a smallest effect size of practical or theoretical interest. In these cases, the use of conventions appears to be the most feasible option. A convention may apply for psychology in general or be specific to a subdiscipline or area of study. An effect equivalent to a correlation of $r = .10$ appears to be a reasonable general interim effect size for study design if the researchers want a good test of a hypothesis and have the needed resources. This tentative suggestion is subject to revision as more experience is gained with confirmatory research. After the results of one or more well-powered studies without researcher flexibility are available, methods noted above for planning research based on previous studies could be used if a smallest effect size of practical or theoretical interest remains elusive.

Sample Size Based on Available Resources

In practice, the sample size for a planned study may be determined by the available resources that the investigators are willing to commit to the project. With online data collection

and big data databases (Vezzoli & Zogmaister, 2023), available resources can sometimes result in studies with extremely large sample sizes, as well as the historical problem of small sample sizes.

This can include cases when a large sample size is used that is apparently considered to provide an adequate test of a hypothesis. For example, Protzko et al. (2020) used samples sizes of about 1500 to investigate 16 new research findings. Although a smallest effect size of practical or theoretical interest may not be explicitly discussed in such cases, the large sample sizes allow falsifiable research. Power analyses as discussed below reveal the effect sizes that are treated as not of interest in these studies.

For the traditional problem of low power, Maxwell, Kelley, and Rausch (2008) noted that the effect size in a power analysis can be selected to fit a predetermined sample size, rather than to determine the sample size. Similarly, a power analysis based on a mean effect size from exploratory research and a design power of .80 can often be used to justify a relatively small sample size. The problem or deceptiveness in these cases is presenting the research as based on a power analysis rather than openly acknowledging that the sample size was based on resources.

Underpowered confirmatory research can be useful if handled transparently. The study preregistration should openly state that the planned sample size was set based on available resources or time limitations. This decision should not be obscured by a superficial, unrealistic power analysis. The power curve for the planned analysis and sample size would reveal the effect sizes that are actually being investigated. The preregistration would state that an inconclusive study outcome may occur due to the limited sample size. The preregistration would also provide adequate detail to eliminate exploratory flexibility that could bias the outcome. At a minimum, such a study could contribute to a meta-analysis of similarly preregistered studies, preferably a prospective meta-analysis (discussed in the last section below).

Underpowered confirmatory research can also occur in other situations. For example, a study exploring covariates of an effect may have small sample size, but the effect may be obtained without researcher flexibility. The flexible exploratory component would pertain to the covariates. This is another case that may be suitable for inclusion in a meta-analysis of preregistered studies without researcher flexibility.

Evaluate the Inferential Error Rates

Many statistical problems and debates will be automatically resolved when researchers recognize that the validity and reliability of a planned confirmatory statistical analysis need to be evaluated, similar to the validity and reliability of measurement instruments. Mayo (2018) calls this the *performance* of a planned statistical analysis. The usefulness of any mathematical model, particularly statistical models, must be based on practical evaluations, not idealistic assumptions.

The performance of a statistical analysis can be evaluated by examining the expected rates of incorrect inferences. These rates are obtained by direct simulations or mathematical models that give the distribution of possible outcomes if the study were repeated many times. This pre-data evaluation is a fundamental step for planning good confirmatory research and provides confidence in the study.

If the inferential error rates are not evaluated, the researchers usually will not know if the planned statistical criteria are feasible for the study—as was common practice with pre-replication-crisis methodology. This practice typically results in under-powered studies and contributed substantially to the replication crisis (Nelson et al., 2018).

A recently attempted multi-lab replication of one of Bem's (2011) studies of precognition shows how this plays out with study preregistration. Maier et al. (2020) stated in the preregistration that the confirmatory study would collect data until a Bayes factor of 10 or 1/10 was reached. They apparently did not do simulations to evaluate the expected sample size to meet these criteria or the expected inferential error rates. Contrary to the preregistration, they stopped the study after 2004 subjects without reaching one of the pre-specified Bayes factor criteria. With this protocol deviation, the main study conclusion that the outcome was moderate support for the null model was an unplanned post hoc inference, as is common for exploratory research.

For comparison, in designing a similar confirmatory study of Bem's research, Kekecs et al. (2023) conducted a thorough evaluation of the operating characteristics that provides high confidence in the planned analysis and in the experimenters' understanding of the analysis. The final study design developed from the simulations had >.95 probability of reaching the Bayes factor criterion (25) for the null model if the null model is true and >.95 probability of reaching

the criterion (1/25) for the alternative model if it is true. The study outcome for the 2115 subjects was very strong evidence for the null model.

Power Curve or Operating Characteristics

The evaluation of inferential error rates or power analyses can include three stages. Not all stages will be needed for every study. These stages apply to Bayesian analyses as well as to frequentist analyses.

The first stage is to determine the tentative sample size. For falsifiable research, this can be based on achieving high power for a prespecified minimum effect size for study design. Recommended magnitudes for power and related parameters are discussed below. The sample size evaluation may be done with mathematical models such as the G*Power software for power analysis (Buchner et al., 2021), or with simulations. Simulations are relatively easy to conduct with current technology (Carsey & Harden, 2014) and will usually be required for Bayesian analyses. Simulations are particularly useful for sequential analyses to determine the expected sample size and variation in sample size, as well as power and inferential error rates. The Bayesian power analyses recommended by Kruschke (2015) and similar "design analyses" described by Schönbrodt and Wagenmakers (2018) are intended for estimating sample size for study planning. Alternatively, if the sample size is set based on available resource or other limitations, analyses at this stage will be skipped.

The second stage is to evaluate the performance of the planned analysis over the range of possible true values for the effect. This provides a useful understanding of which effect sizes can be reliably detected and which cannot. The inferential error rates or proportion of correct inferences are determined when the planned analyses and tentative samples size or stopping criteria for sequential analyses are applied to hypothetical true effects that range from no effect to a very strong effect. The resulting power curve or operating characteristics is a table or graph with proportion of correct inferences by hypothetical true effect size. For falsifiable research, the inferential error rates for inferences that the alternative hypothesis is false would be evaluated as well as for inferences that the null hypothesis is false. Here too, power curves can be based on either mathematical models or on simulations. The G*Power software has an option "X -Y plot for a range of values" than can generate a power curve graph (Buchner et al., 2021). Simulations would be needed for sequential analyses.

For Bayesian analyses, these performance evaluations are described as the operating characteristics. They are expected for confirmatory medical research seeking approval from the U.S. Food and Drug Administration (2010). That is the best discussion of Bayesian analysis for confirmatory research that I have found. Bayesian prior probability distributions can have practical implications that are not easy to understand, including substantial biases that make a hypothesis test have unintended and unwanted properties (Gu et al., 2016; Kruschke, 2015; Simonsohn, 2015; Tendeiro & Kiers, 2019). The operating characteristics make the practical implications of prior probability distributions clear and reveal potential biases. The research plan developed by Kekecs, et al. (2020) used this method when planning a sequential Bayesian analysis and is a useful model for effective statistical planning. Bayesian analysis will remain of limited value for confirmatory research until software for developing operating characteristics is more readily available.

The best practice for evaluating the performance of a planned confirmatory analysis is: (a) before data collection begins, develop and validate the data analysis scripts or programs that will be used for the final analysis, (b) use the scripts or programs in simulations that evaluate the performance of the planned analysis, and (c) include the scripts or programs in the preregistration for the study. These steps should be possible with appropriate previous exploratory research. This process clearly establishes that the planned analyses were prospectively developed for the confirmatory research. Deviations from the preregistered analyses should be described and explained as protocol deviations. Using the scripts or programs in simulations for evaluating the performance of the planned analyses is also a useful step in validating the software code.

The third stage is to conduct sensitivity analyses that evaluate how the performance (inferential error rates) of the planned analysis is affected if the assumptions for the power analysis are changed. This provides an even better understanding of the strengths and weaknesses of the planned analysis. Sensitivity analyses are particularly valuable for studies with complex statistical models and/or studies that are controversial or have significant practical implications. For example, sensitivity analyses may evaluate how the expected error rates are affected by changes in the assumptions for random effects, covariates, dropouts, and ceiling effects. The final planned sample size may be set based on the uncertainties revealed by the

sensitivity analyses. Here too, a sensitivity analysis for falsifiable research would evaluate errors in inferences that the alternative hypothesis is false as well as that the null hypothesis is false. Sensitivity analyses will often be done with simulations and may not be needed for studies with simple analyses.

Power analysis can incorporate adjustment for multiple analyses. As always, the inferential error rates are estimated when a certain condition is hypothetically assumed to be true. The evaluations cover the range of possible true effects. For example, if adjustment for four analyses is planned, the expected inferential error rates can be estimated when the null models are assumed to be true for all four analyses. Similarly, evaluations can be done with the assumption that the alternative models are true for all four analyses. The possibilities that the alternative models are true for one, two, or three analyses and the null models for the others can also be evaluated. Sufficient evaluations should be done to provide a good understanding of the performance of the planned analyses, but not every possible combination of hypothetically true hypotheses needs to be evaluated. Additional discussion of multiple analyses is given below, including the pivotal question of when adjustment for multiple analyses is needed.

The evaluation of inferential error rates provides confidence in the planned analysis and is relevant before and after data have been collected. Some advocates of Bayesian methods argue that the evaluations of inferential error rates for Bayesian analyses are useful for study design but have no role after the data are collected. That is similar to arguing that the validity and reliability of a measurement instrument are irrelevant after a measurement has been made. The evaluations of the performance of a measurement instrument or a statistical analysis provide confidence in the results both before and after the data have been collected.

Frequentist Evidence that the Alternative Hypothesis is False

The same logic that is used to reject the null hypothesis can be used to reject the alternative hypothesis. The logic for rejecting the null hypothesis starts by developing a statistical model for no effect. If a result as extreme as the observed data would be rare if this null model is true, the data are interpreted as evidence that the null model is not applicable and an effect occurred. Typically, a probability of less than .05 is considered evidence that the null model is not applicable.

Similarly, a statistical model can be developed for the alternative hypothesis that the effect is a certain size. If a result as extreme as the observed data would be rare if the alternative model is true, the data can be interpreted as evidence that the alternative model is not applicable for the conditions of the study (Cohen, 1988, pp. 16-17; Coenen & Smits, 2022; Mayo, 2018, pp. 339-342; Stanley et al., 2022). For falsifiable research, the predicted effect for the alternative model is the minimum effect size for study design.

Note that this inference does not conclude that the null model of no effect is true, only that any effect is smaller than the effect size in the statistical model for the alternative hypothesis. This strategy focuses on inferences about the alternative model, not inferences about the null model.

The statistical model for the alternative hypothesis in a power analysis can be used to implement this strategy. Notably, a study with a power of .95 can reject the alternative hypothesis at the .05 level and is symmetric with rejecting the null hypothesis at the .05 level.

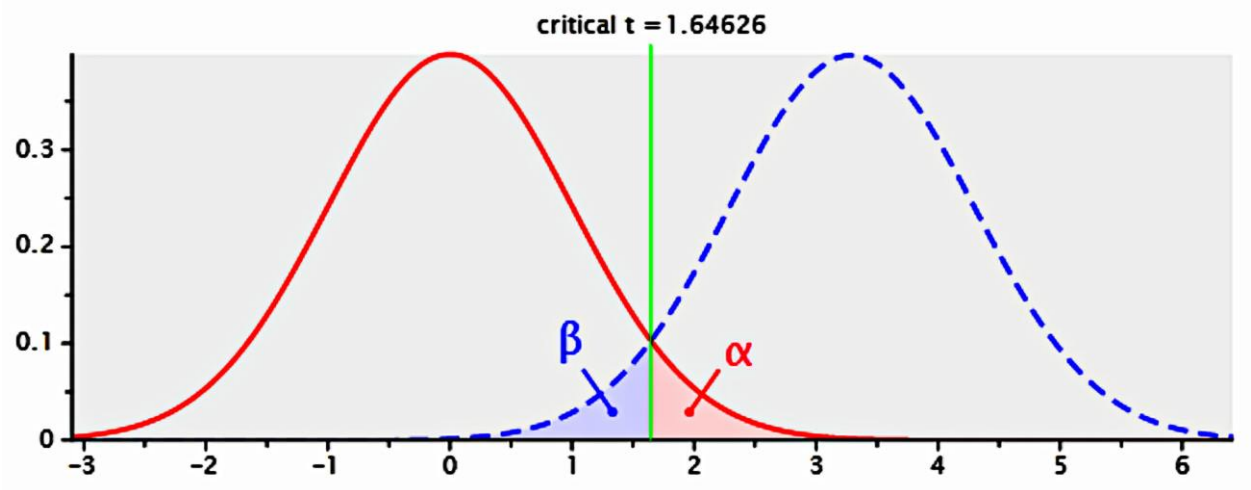
The basic principles can be seen with a simple t-test comparing the means of two samples. Figure 1 displays a power analysis showing the distribution for the null model with the significance level or alpha set to .05 and the distribution for the alternative model with a power of .95. Alpha is the probability of incorrectly rejecting the null model if it is true—a type I or false-positive error. Beta is the probability of incorrectly rejecting the alternative model if it is true—a type II or false-negative error. Power is one minus beta. For a power of .95, beta is .05 and is equal to alpha.

With power of .95, a nonsignificant outcome can be interpreted as rejecting the alternative model in the same way that a significant outcome is interpreted as rejecting the null model. Stanley, Doucouliagos, and Ioannidis (2022) applied this analysis when they compared initial studies with subsequent preregistered multi-lab studies. They referred to it as the “Cohen null” test.

Rejecting the alternative model is equivalent to switching the null and alternative hypotheses, and rejecting the new null hypothesis that the true effect size is the size specified in the original power analysis. However, keeping track of switched null and alternative hypotheses can become confusing. Describing the situation in context of rejecting the alternative hypothesis may be more straightforward for most statistical users.

Figure 1.

Power Analysis Models for Alpha = .05 and Power = .95



Note. Distributions for the null model (left, solid line) and alternative model (right, dashed line) for a two-sample t-test with d = .2, α = .05 and β = .05 for power = 1 – β = .95. An outcome to the right of the vertical line is significant and to the left is nonsignificant. β is the probability of incorrectly rejecting the alternative model when it is true and α is the probability of incorrectly rejecting the null model when it is true. The plot is from the G*Power program (Buchner et al., 2021).

A *p* value for rejecting the alternative hypothesis can be obtained from the statistical model for the alternative hypothesis. This is particularly useful if the power is less than .95. For a simple t-test, the optimal statistical model for the alternative hypothesis is a noncentral t distribution, which is applicable when the assumed true effect in the statistical model is not zero. Noncentral distributions are typically used in power analyses. The manual and related publications for the G*Power software for power analysis provide useful information about the statistical models for alternative hypotheses (Buchner et al., 2021; Faul et al., 2007). Coenen and Smits (2022) describe the mechanics for analyses with noncentral distributions, including more complex models than the simple t-test discussed here. Statistical software such as R have functions for the noncentral distributions.

A *location-shift t-test* can be a useful approximation to a noncentral t-test. The minimum effect size for study design is subtracted from the mean observed effect to shift the analysis to zero if the minimum effect size for study design is true. The usual central t-test (based on a true effect size of zero) is then applied to evaluate whether the shifted data are different from zero,

which translates to different from the minimum effect size for study design. For the t.test function in the R software, the *mu* parameter implements a location-shift t-test. The *mu* parameter would be set equal to the minimum effect size for study design. If the minimum effect for study design is positive and a smaller effect is evidence that the alternative hypothesis is not true, the R t.test parameter *alternative* would be set to "less." The output of the t.test function nicely shifts the mean and confidence interval back to the starting point specified with *mu*. The data and the *mu* parameter must have the same scale, either standardized (converted to *d* values) or unstandardized (raw data). Location-shift t-tests are commonly used in equivalence tests, which were developed for two-sided tests and are discussed below.

The location-shift t-test and noncentral t-test become identical as sample size increases (Cumming & Finch, 2001). However, location-shift t-tests tend to be conservative compared to noncentral t-tests for smaller sample sizes. One suggested guideline for medical bioequivalence research was that an equivalence test using location-shift t-tests has power similar to other methods for studies with power of .80 or higher (Meredith & Heise, 1996). More extensive simulation studies would be useful, particularly with psychological data. The location-shift model can be adapted to tests with unequal variances and to nonparametric tests.

A power analysis for rejecting the alternative hypothesis can be obtained by switching the magnitude of alpha and beta in the usual power analysis of the null hypothesis. For a case with alpha of .05 and power .90, beta is .10. The sample size for rejecting the alternative hypothesis with the same power is obtained by setting alpha to .10 and beta to .05. For most statistical tests, the sample size for alpha of .05 and beta of .10 is the same or only slightly different than for alpha of .10 and beta of .05. This can be easily seen by doing example power analyses with the G*Power software. Thus, the magnitudes of alpha and beta can be switched and the alternative model can be rejected at the .05 level with .90 probability of correctly inferring that the alternative model is not true if the null model is true. For confirmatory research, the safest practice when alpha and beta are not equal is to use the largest sample size when the magnitudes of alpha and beta are switched in the power analysis.

A study outcome can reject both the null model and the alternative model. This is evidence that the effect is not zero, but is smaller than the minimum effect size for study design.

If neither the null model nor the alternative model is rejected, the outcome is inconclusive and indicates low power.

Confidence Intervals

For falsifiable research, confidence intervals generally provide more intuitive and more informative inferences than *p* values. Confidence intervals can provide additional information that enhances or limits an inference and are less prone to misleading interpretations. The same power analysis used for a hypothesis test with *p* values can be conducted as a convenient way to set the sample size to obtain confidence intervals that provide useful information about the hypotheses of interest. However, the more refined inferences from confidence intervals can be incorporated by doing simulations that utilize confidence intervals.

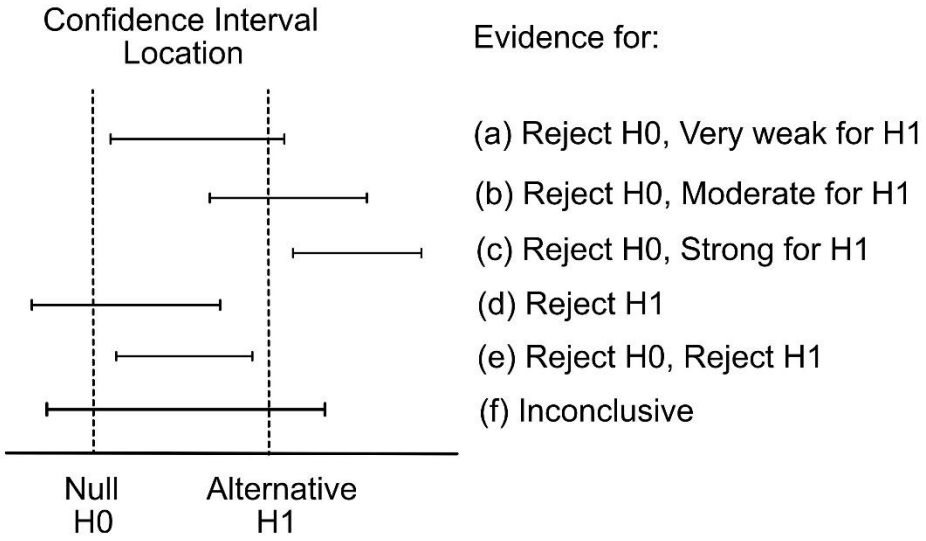
An evaluation of confidence intervals is based on the location of the confidence interval relative to the minimum effect size for study design and to zero or the null model. The interpretations are described below for the example of a one-sided prediction of a positive effect. For confidence intervals equivalent to one-sided hypothesis tests with alpha of .05, 90% two-sided confidence intervals would be used (with 5% on each end) rather than the usual 95% confidence intervals (with 2.5% on each end). These are confidence intervals for the observed mean, not confidence intervals for an individual observation. The inferences described below are displayed in Figure 2.

Evidence that the null hypothesis is false is obtained if the lower end of the confidence interval is greater than zero. However, rejecting the null model is not necessarily good evidence that the alternative model is true. Evidence for the alternative model is based on whether most or all of the confidence interval is above the minimum effect size for study design. If the entire confidence interval is above the minimum effect size for study design, that is strong evidence that the alternative hypothesis is true. If this is made a requirement for an inference, it becomes what is now known as a minimum-effect test (Murphy et al., 2014). This test is particularly useful when the minimum effect is a hard boundary with practical consequences. For most research, the minimum effect for study design is a soft boundary with researchers interested in outcomes near the boundary that may have confidence intervals that extend below the boundary. Specific criteria could be set for moderate evidence that the alternative hypothesis is true, such as

that the mean or perhaps the lower 70% one-sided confidence bound must be above the minimum effect size for study design.

Figure 2.

Inferences from Confidence Intervals



Note. The vertical dashed lines indicate the expected effect sizes for the null model (H0) and for the minimum effect size for study design for the alternative model (H1).

Evidence that the alternative hypothesis is false is obtained if the upper bound of the confidence interval is less than the minimum effect size for study design.

Evidence that both the null model and the alternative model are false is obtained if the lower end of the confidence interval is above zero and the upper end is below the minimum effect size for study design. This is evidence that the effect is greater than zero and smaller than the minimum effect size for study design.

The outcome is inconclusive if the lower bound of the confidence interval is below zero and the upper bound is above the minimum effect size for study design. This indicates a wide confidence interval and low power for the study.

For underpowered confirmatory research, an investigator may choose to report only effect sizes and confidence intervals without any type of inference about hypotheses. Of course, this decision must be preregistered and not chosen after looking at the results.

One decision that merits some thought is whether to report the effect sizes and confidence intervals in unstandardized units or in standardized units such as correlation coefficients and Cohen's *d*. With standardized effect size measures, the magnitude of the effect is divided by the standard deviation, which converts the effect to units of standard deviation and is essentially a signal-to-noise ratio. This ratio loses information about the magnitude and variability of the effect. For example, an effect with a large magnitude and large standard deviation can have the same standardized effect size as an effect with a small magnitude and small standard deviation. Distinguishing these cases can be useful, particularly in applied research. At the same time, a signal-to-noise ratio or standardized effect size can indicate the practical usefulness of an effect and is used in power analyses. Standardized effect sizes are used in meta-analyses and widely reported in psychological research because that allows direct comparison with findings from meta-analyses. However, some writers advocate reporting unstandardized effects because that provides more complete information and is less prone to obscuring inconsistencies in the data (Baguley, 2009; Lenth, 2001; Pek & Flora, 2018). Given the trade-offs, one obvious option is to report both, with one possibly being included in supplemental materials.

A related point is that confidence intervals for standardized effect size measures require special calculations that are a still emerging methodology, particularly for repeated measures (Cousineau & Goulet-Pelletier, 2021; Goulet-Pelletier & Cousineau, 2018). If confidence intervals for standardized effect size measures will be used for inferences about hypotheses, the investigators would be well-advised to become familiar with the current state of the art and the specific algorithms in the software they are using. Using those algorithms in simulations for power analyses would provide greater confidence in the inferences.

Magnitude of Power

In general, for falsifiable confirmatory research, a power of ≥.95 can be considered very good and .90 can be considered good. Power of .80 has historically been considered adequate for setting sample size. Cohen (1965) originally recommended a power of .80 in the 1960s when data collection and often data analyses were done with paper and pencil. He believed that studies

with power much greater than .80 would usually be infeasible or not worth the effort. In addition, he thought that a false-positive (Type I) error was about four times more serious than a false-negative (Type II) error. With a power of .80 a statistical analysis has .20 probability of failing to detect a true effect and this is four times larger than alpha of 05.

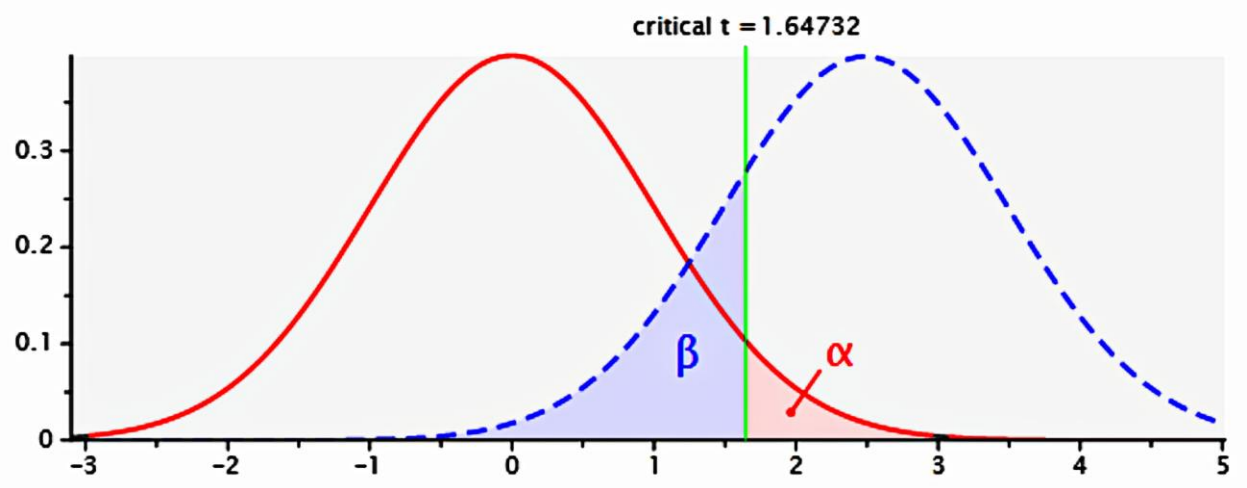
The bias that occurs with a power of .80 is shown in Figure 3. The Y axis is the probability density that indicates how likely an outcome on the X axis is for the null model and for the alternative model. Note that an area of inversion occurs to the left of the critical value, before the curves cross. An outcome in that area is treated as consistent with the null model, but the outcome is actually more likely and more consistent with the alternative model. This bias is implicitly deemed acceptable for power of .80 because false-negative errors are considered less objectionable than false-positive errors. An analogous inversion to the right of the critical value will occur if the power is high and beta is smaller than alpha. In that case, some outcomes that would be more likely with the null model will be treated as supporting the alternative model. This implicitly treats false-positive errors as more acceptable than false-negative errors. However, with power greater than .95, an outcome in the area of inversion (just below $p = .05$) is unlikely if either the null or the alternative model is true. As shown in Figure 1, the inferences are always consistent with the models if alpha and beta are equal.

For confirmatory research that is theoretically oriented rather than applied, false-positive and false-negative errors usually can be considered about equally undesirable. This is implemented by setting alpha and beta equal when power is above .95. However, setting alpha and beta equal may not be optimal for applied research that has different consequences for false-positive and false-negative errors.

As examples of the needed sample sizes, the overall sample sizes for one-sided tests with Cohen's *d* of .20 for a two-sample t-test with equal groups, alpha of .05, and powers of .80, .90, and .95 are 620, 858, and 1084. For Cohen's *d* of .41 the corresponding sample sizes are 150, 206, and 260. The sample sizes for correlation coefficients of .10 and .20 are similar to those for Cohen's *d* of .20 and .41. Most of the Registered Replication Reports published in *Advances in Methods and Practices in Psychological Science* have sample sizes of over 1000, and some have sample sizes of several thousand.

Figure 3.

Power Analysis Models for Alpha = .05 and Power = .80



Note. Distributions for the null model (left, solid line) and alternative model (right, dashed line) for a two-sample t-test with d = .2, α = .05 and β = .20 for power = 1 – β = .80. An outcome to the right of the vertical line is significant and to the left is nonsignificant. The area to the left of the critical value before the curves cross is treated as consistent with the null model, but those outcomes are more likely with the alternative model.

A project that implemented multiple confirmatory studies with sample sizes of about 1500 for 16 new experimental findings found reliable confirmations consistent with the power analyses (Protzko et al., 2020). For each of the 16 effects, the laboratory that developed the new effect with exploratory research conducted a preregistered confirmatory study that was followed by three preregistered confirmatory studies by other laboratories. The research was limited to effects that could be investigated with tests of the difference between two independent samples. The authors note that the high reliability in these cases may not apply for more complicated designs.

As power approaches 1.0, p = .05 becomes less meaningful. With power above .95, an outcome rejecting the null hypothesis with p near .05 may be false-positive, and the alternative hypothesis may also be rejected with about the same p value. The G*Power software can be used to set alpha and beta equal, which improves the inferences. However, confidence intervals will be far more revealing about what inferences are actually supported.

Multiple Analyses

Opinions about when corrections for multiple analyses are needed vary greatly (García-Pérez, 2023; Rubin, 2021). The literature is large and the conflicting opinions are sometimes strongly held. Rubin (2021) provides an overview with many references. The ideas presented here are consistent with the arguments of Rubin (2021) and with many writers preceding him, but have a simpler rationale and terminology here.

For confirmatory research, correction for multiple analyses is necessary when any one of several analyses or effects would be considered evidence for a higher-level or more general hypothesis. For example, a medical intervention may be evaluated by examining several different health measures. If a positive outcome for any one of the health measures would be interpreted as evidence that the intervention is effective, correction for multiple analyses is needed. Each of the health measures is a component of the higher-level hypothesis that the intervention is effective. Evidence for the higher-level hypothesis can be obtained by selecting results from the individual components. The multiple analyses may be handled by adjusting the result for each component, by combining the components into a scale with a single outcome, or by a multivariate analysis of the components. Correction for multiple analyses can be considered optional with the greater flexibility of exploratory research.

Correction for multiple analyses is increasingly recognized as unnecessary when each analysis is for an individual confirmatory hypothesis that stands alone and is not a component of a higher-level hypothesis in the study (Rubin, 2021). For example, a study of lifestyle and health could include predictions that smoking will be associated with lung cancer and that consumption of saturated fats will be associated with cardiovascular disease. Strong resistance can be expected for arguments that the evaluation of lung cancer must be adjusted for the investigation of cardiovascular disease because they happened to be investigated in the same study rather than in different studies. The possibility of a false-positive result for lung cancer is the same as for a study that investigated only lung cancer.

A researcher planning a study with six independent standalone confirmatory hypotheses may have discomfort with the fact that the probability of one or more false-positive results is $1 - .95^6 = .26$ (assuming the null hypotheses are true and alpha = .05). The researcher should consider whether the discomfort equally applies if the six hypotheses are tested in separate

studies, which would give the same overall .26 error rate. The assumption that the hypotheses are independent in this example is for simplicity in calculating probabilities. Standalone hypotheses need not be independent as long as they are not components of a higher-level hypothesis.

The assumption that correction for multiple analyses is widely needed may have originated with the pre-replication-crisis tradition of focusing almost exclusively on significant *p* values. A study was considered successful to the extent that one or more of the analyses produced a significant *p* value. Publication was contingent upon significant *p* values (Ritchie, 2020). Adjustment for all the analyses in a study could be reasonably recommended with these priorities. However, for current confirmatory research, those priorities no longer apply. Each standalone hypothesis or prediction can be evaluated on its own merits.

Researchers may have differing opinions about whether a higher-level hypothesis is applicable, particularly in controversial areas of research. For example, a proponent of precognition might plan a study with five analyses that look at different possible manifestations of precognition. The proponent may accept that precognition occurs and consider each of the five hypotheses as standing alone and confirming findings in previous experiments. However, a skeptic of precognition would view these as five components of the higher-level hypothesis that precognition occurs, and would expect correction for multiple analyses.

The Costs of Correction for Multiple Analyses

Correction for multiple analyses comes at a high price in terms of falsifiable research. The case of two planned independent confirmatory analyses shows the trade-offs with multiple analyses. Consider a case with alpha and beta set to .05 for both analyses without adjustment for multiple analyses. For each of the two tests, if the null model is true, the probability that the analysis will give the correct inference is .95. If the null model is true for both tests, the probability that both analyses will give correct inferences is .95 x .95 = .90. The combined error rate is 1 – .90 = .10 rather than .05. The typical Bonferroni correction changes alpha from .05 to .025 and gives .975 x .975 = .95, for a combined error rate of .05.

Similarly, a combined power of .95 for the two tests for falsifiable research would be obtained by increasing the sample size to produce a Bonferroni correction for beta as well as for alpha. If the alternative model is true for both tests, the probability that both tests will obtain correct inferences without adjustment for multiple analyses is .95 x .95 = .90. The power for each

test must be increased to .975 to obtain a combined power of .95. The Bonferroni adjusted alpha and Bonferroni adjusted power would be used to determine sample size. Most discussions of multiple analyses focus on rejecting the null hypotheses without considering falsifiable research or the performance of the planned statistical analysis over the range of possible true hypotheses.

More extensive discussion of the methods and controversies for handling multiple analyses is beyond the scope of the present paper. However, a few observations may be helpful. First, given the current differences of opinion, detailed preregistration of the planned confirmatory hypotheses and analyses is essential. This allows readers to determine for themselves whether adjustment is needed. In a related point, unadjusted *p* values and confidence intervals should be reported even if adjusted values are also reported. Over time, the unadjusted values will be more useful as the results are compared with the results from other studies. Second, adequately powered falsifiable confirmatory research with correction for multiple planned analyses may be possible when large databases are available. Confidence intervals and *p*-values would be adjusted for multiple analyses. Third, not all confirmatory studies are or need to be falsifiable. Research with correction for multiple analyses may appropriately be considered as underpowered confirmatory research or perhaps confirmatory research with unknown power. In these cases, evidence rejecting the alternative model may sometimes be obtained, but that is due more to happenstance than to careful planning.

Related Frequentist Ideas

Inferences based on the statistical model for the alternative hypothesis are also central components of statistical philosopher Deborah Mayo's (2018) *severe testing* and Coenen and Smits's (2022) subsequent *strong-form testing*. However, their logic focuses on counterfactual reasoning. For example, the severity interpretation of a lower .95 one-sided confidence interval is that if the true effect size were less than the lower .95 confidence limit, the probability would be ≥.95 that the study would have found a mean effect size that was less than the mean effect size that was actually found (Mayo, 2018, p. 195). Although counterfactual reasoning may be technically correct in terms of certain philosophies about probability, the arguments are sometimes difficult to follow for those of us who are primarily interested in practical application of statistical methods. The goals of severe testing are a valuable step forward for research methodology, but the logic and methods will probably need to be made more intuitive and

directly useful before they will be widely adopted. The writings about severe testing appear to focus on confirmatory research, but the distinction between the exploratory and confirmatory stages of research is not explicitly addressed.

Using equivalence tests to provide evidence that the data are consistent with the null hypothesis has been suggested as a strategy for falsifiable psychological research (Lakens et al., 2018; Maxwell et al., 2015). Equivalence tests were developed to provide evidence that two medical treatments have equivalent effects. As with the evaluation of alternative hypotheses described above, equivalence tests are based on rejecting a statistical model that the two treatments are different by at least a specified smallest effect size.

A basic equivalence test is two-sided and is implemented with two one-sided tests (TOST), one for the upper side and one for the lower side. To establish equivalence, both tests must reject the hypotheses that the treatment differences are more extreme than the specified smallest effect. Adjustment for multiple analyses is not needed because rejection by both tests is required, not just one or the other, and only one of the tests can be an incorrect rejection (Meyners, 2012). However, a power analysis for TOST does require adjustment because either test can incorrectly accept the hypothesis that the treatments are different (Julious, 2004). A power of .95 for each test is needed to have a combined power of .90. The result of TOST is the same as the 90% (not 95%) two-sided confidence interval entirely contained within the range of effect sizes considered as equivalent.

The conceptual framework and software options for equivalence tests are focused on two-sided tests. For one-sided tests, location-shift t-tests described above can be easily implemented with the usual software for t-tests, and can be considered one-sided equivalence tests.

Bayesian Evidence that the Alternative Hypothesis is False

Bayesian methods can also provide evidence that an experimental hypothesis is false. Credible intervals can be used for inferences similar to confidence intervals. The most common method is the Bayes factor that compares and quantifies the extent to which the observed data would be expected with an alternative model versus with the null model (Dienes, 2014; Mulder & Wagenmakers, 2016). Although extensively used conventions have not yet been developed, the criteria are often applied that data that are three to five times more likely to occur with the

null model than with the alternative model are considered moderate or substantial evidence that the alternative model is not true, and data that are ten times more likely are strong evidence.

High odds in favor of a model are evidence that one model is more consistent with the data than is the other model, but Bayes factor odds are not direct evidence that a model is true (Tendeiro & Kiers, 2019). The Bayes factor identifies the better of two models, but both models could be invalid. Thus, Bayes factor odds in favor of the null model are evidence that the alternative model is not applicable, but not necessarily evidence that the null model is true.

Here too, examination of the operating characteristics for the analysis may reveal that the analysis does not perform well for certain true effect sizes. Careful evaluation of the operating characteristics during study design can provide confidence that a design and analysis provide reliable inferences that one of the models is valid. The optimal power of ≥.95 and good power of .90 are applicable for confirmatory Bayesian analyses as well as for frequentist analyses.

As always with Bayesian methods, the inferences are contingent upon the prior probability distributions that were selected. Kruschke (2015) argues that Bayes factors can be extremely sensitive to the choice of prior probability distribution, and he presents the Region of Practical Equivalence as an alternative strategy. ROPE is an estimation approach analogous to drawing inferences from confidence intervals in frequentist statistics. Evaluation of operating characteristics is particularly important for analyses that are sensitive to prior probability distributions, but is needed for any planned confirmatory inferences, including those with ROPE. Tendeiro and Kiers (2019) discuss various limitations and misinterpretations associated with Bayes factors.

Preregister the Statistical Methods and Inference Criteria

To the maximum extent possible, confirmatory research should be preregistered on a public study registry. The best practice for confirmatory research when possible is a Registered Report (Chambers & Tzavella, 2022). The study plans are subject to peer review before data collection begins. The pre-approved plans should be posted on a public study registry like any other study. As described by Chambers and Tzavella (2022), Registered Reports are a subset of preregistered studies, not an alternative to preregistration.

When a planned research project involves an exploratory phase followed by a confirmatory phase, the confirmatory phase should be preregistered after the exploratory phase has been completed and the details for the confirmatory predictions have been developed. This includes cases when a large database is split for exploratory and confirmatory analyses, as well as when new data are collected in sequential phases of a project.

The preregistration for a study should specify which analyses are confirmatory and which are exploratory, and the statistical methods and inference criteria that will be used for the confirmatory analyses (Wicherts et al., 2016). The inference criteria are the actual numerical criteria for inference, not just the type of statistical test. These criteria include whether the test is one-sided or two-sided, and the magnitude of the *p* value, confidence interval, odds, or other statistical parameter that will be considered evidence for an inference. For Bayesian methods, the selected prior probability distributions should also be included in the preregistration. The planned handling of multiple analyses should also be specified, if applicable.

The best practice is to develop and validate the scripts or programs for data processing and analyses prior to starting data collection, and include them as part of the preregistration. If researchers need or want to look at the confirmatory data before developing the analysis scripts or programs, the analyses would probably be more appropriately classified as exploratory in most cases.

For falsifiable confirmatory research, the preregistered inference criteria should include (a) criteria for evidence the alternative hypothesis is true, (b) criteria for evidence the alternative hypothesis is false, and (c) criteria for outcomes that will be inconclusive. The usual statistical methods can be used as evidence that the alternative hypothesis is true. The inference criteria can be based on confidence intervals, credible intervals, *p* values, or Bayesian odds.

Similar methods can be used to pre-specify the criteria for evidence that the alternative hypothesis is false. Here too, inferences can be based on confidence intervals, credible intervals, *p* values, or Bayesian odds.

Inconclusive outcomes should be addressed in the preregistration if outcomes that do not meet the above criteria are possible. Inconclusive outcomes are most likely if the studies do not have high power. The preregistration in this situation should specify the outcomes that will be

considered inconclusive. Similarly, if the study is underpowered and the investigators plan not to make inferences about a hypothesis, that should be clearly stated in the preregistration.

It is also possible that the confidence interval and associated *p* values will reject both the null and the alternative models. This can occur if the study has high power and the true effect size is larger than zero and smaller than the minimum effect size for study design. The preregistration should address what inferences will be made in this situation. A good power curve evaluation can provide insights about this possibility.

An optimal preregistration will provide inference criteria for all reasonably foreseeable outcomes. One common case is that a maximum sample size is specified for a sequential analysis. Any inferences that will be made if the maximum is reached should be prespecified. For example, an analysis plan could state that data collection will stop when a Bayes Factor of 10 or 1/10 is reached or the sample size is 2000. The preregistration could also state that with a sample size of 2000, Bayes Factor odds between 5 and 10 or between 1/5 and 1/10 will be considered substantial support for the alternative or null models. If the desired sample size may not be reached for other reasons (such as time or resource limitations), the inferences that will be made with smaller sample sizes should be specified in the preregistration. If unanticipated outcomes are likely, the researchers should consider whether the study is actually exploratory.

Inferences based on study outcomes that are outside the preregistered possibilities should be openly described as unplanned or post hoc when reporting the study. The credibility of such an inference will depend on the degree of researcher flexibility in the occurrence and interpretation of the outcome. Outcomes outside the preregistered possibilities are a form of protocol deviation that should be explained and evaluated like other significant protocol deviations. A few protocol deviations can be expected for a large study. Researchers can also expect that undisclosed significant protocol deviations will eventually be revealed when the study report is compared with the preregistration (Claesen et al., 2021; Goldacre, et al., 2019).

Although the analyses described here involve two hypothesis tests (one for the null and one for the alternative hypotheses), correction for multiple analyses may generally be unnecessary. This is not the problem situation when researchers have multiple opportunities to declare evidence for an effect. At the same time, if the usual 95% two-sided confidence interval

is used for inferences, that will automatically adjust for two analyses because the upper and lower bounds are each the same as for a one-sided test with alpha of .025.

Final Thoughts

The methods described above may be considered idealistic. However, they are achievable in many situations, and they are a useful frame of reference for considering the strengths and limitations of statistically-based research in general. Three additional topics related to confirmatory research are discussed below.

Making Preregistrations Useful

The self-registration processes at sites such as Open Science Framework (OSF) tend to have incomplete preregistrations that do not completely eliminate researcher flexibility (Bakker at al., 2020). Bakker et al. found that structured preregistration forms were more complete than unstructured preregistrations, but "neither performed impressively." With current practices, the initial version of a preregistration can be presumed to be incomplete. A study registry that provides a review of the submitted preregistrations for completeness would be a significant improvement, but is rare for available study registries (Watt & Kennedy, 2015). As an alternative, researchers could ask a colleague to review a draft preregistration for completeness. Someone who is knowledgeable of methodology, detail oriented, and does not have knowledge and assumptions about the project would be best.

The preregistration form of the American Psychological Association is one of the most advanced and one of the few to encourage specification of a smallest effect size of interest (American Psychological Association, 2021). However, the form is still basically a checklist with a few hints that relies on the researcher for completeness.

Study registries should make the preregistrations irreversibly public. Websites such as AsPredicted.org that allow experimenters to make a preregistration public only if the results are favorable defeat the value of preregistration for preventing researcher flexibility and publication bias. Nosek et al. (2018) stated "the website https://aspredicted.org/ provides a simple form for preregistration, but it is not itself a registry because users can keep their completed forms private forever and selectively report preregistrations."

Inconsistency between the report for a study compared to the preregistration occurs to a surprising degree (Claesen et al., 2021; Goldacre, et al., 2019). Preregistration must be treated as a means to detect bias as well as to prevent bias. Prior to or during review for publication, the consistency between the draft report of the study and the preregistration should be carefully evaluated. Of course, researchers should also track these inconsistencies as they are conducting the study and preparing the report.

Retrospective Versus Prospective Meta-Analysis

Retrospective meta-analysis is a type of post hoc analysis with much flexibility in the decisions made by the analysts (Watt & Kennedy, 2017). The replication-crisis lessons about researcher flexibility apply to retrospective meta-analyses as well as to individual studies. The post hoc, flexible nature of retrospective meta-analysis is much closer to exploratory research than to confirmatory research. A retrospective meta-analysis that focuses on or has subsets for preregistered confirmatory studies is a step forward, but is still a post hoc methodology with researcher flexibility.

Prospective meta-analysis is a much-preferred option that is confirmatory in nature (Watt & Kennedy, 2017). One type of prospective meta-analysis is similar to a large multi-lab study. Another option is *registration-based prospective meta-analysis*. With this option, the planned methodology for the meta-analysis is preregistered. The decision to include a particular subsequent study is based on the preregistration for the study before the results for the study are known, and ideally before the study has been conducted. A list of studies that will be included in the meta-analysis is maintained and updated in a public document along with the preregistration for the meta-analysis. Study preregistrations will need to be more complete and more consistent before this methodology can be used optimally, but limited use is possible at present. Obviously, the preregistrations will also need to be publicly available. The best practice for confirmatory research is preregistrations that facilitate potential inclusion in a prospective meta-analysis.

Research Quality Control

The validity of research findings depends on much more than the statistical analyses. Quality control measures for study conduct are among the important factors to consider. Quality control measures include software validation, double checking all data handling, and measures to

prevent experimenter fraud. These are well established research problems that threaten the validity of research findings, but at present are rarely discussed in research reports or methodological papers—which focus instead on statistical methods and results. Notably, it is now well established that peer review and replication are generally not effective at deterring or detecting experimenter fraud, and the frequency of undetected fraud is likely much higher than the frequency of detected fraud (Nelson et al., 2018; Ritchie, 2020; Strobe, Postmes, & Spears, 2012). Quality control topics are discussed with examples and recommendations in Kennedy (2023).

References

American Psychological Association (2021, January). *Preregistration*. https://www.apa.org/pubs/journals/resources/preregistration

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(S1), 262-270. https://doi.org/10.1080/00031305.2018.1543137

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. (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

Anvari, F. & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. Journal of Experimental Social Psychology, *96*, 1-11. https://doi.org/10.1016/j.jesp.2021.104159

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617. http://dx.doi.org/10.1348/000712608X377117

Bakker M., Veldkamp, C.L.S., van Assen, M.A.L.M., Crompvoets, E.A.V., Ong, H.H., Nosek, B.A., et al. (2020) Ensuring the quality and specificity of preregistrations. *PLoS Biology,* 18(12): e3000937.https://doi.org/10.1371/journal.pbio.3000937

Bem D. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. https://psycnet.apa.org/doi/10.1037/a0021524

Bouwmeester, S., Verkoeijen, P. P. J. L., Aczel, B., Barbosa, F., Bègue, L., Brañas-Garza, P., … Wollbrant, C. E. (2017). Registered Replication Report: Rand, Greene, and Nowak (2012). *Perspectives on Psychological Science*, *12*, 527–542. doi: 10.1177/1745691617693624

Buchner, A., Erdfelder, E., Faul, F., & Lang, A-G. (2021). G*Power: Statistical Power Analyses for Mac and Windows [software]. https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

Carsey, T.M., & Harden, J.J. (2014). *Monte carlo simulation and resampling methods for social science*. Los Angeles: Sage.

Center for Open Science (n.d.). Preregister your next study. https://www.cos.io/initiatives/prereg

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40 –48. https://doi.org/10.1177/1745691613513470

Chambers, C.D. & Tzavella, L. The past, present and future of Registered Reports. *Nature Human Behavior*, *6*, 29–42 (2022). https://doi.org/10.1038/s41562-021-01193-7

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8, 211037211037. http://doi.org/10.1098/rsos.211037

Cohen, J. (1965). Some statistical issues in psychological research, in B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

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

Coenen, L. & Smits, T. (2022). Strong-form frequentist testing in communication science: Principles, opportunities, and challenges, *Communication Methods and Measures*, *16*(4), 237-265. https://doi.org/10.1080/19312458.2022.2086690

Cousineau, D. & Goulet-Pelletier, J.-C. (2021) A study of confidence intervals for Cohen's dp in within-subject designs with new proposals. *The Quantitative Methods for Psychology, 17*(1), 51-75. https://doi.org/10.20982/tqmp.17.1.p051

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-573. https://doi.org/10.1177/0013164401614002

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Earp, B.D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social Psychology. *Frontiers in Psychology, 6*, 621. http://dx.doi.org/10.3389/fpsyg.2015.00621

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

Ferguson, C. J., & Heene, M. (2021). Providing a lower-bound estimate for psychology's "crud factor": The case of aggression. *Professional Psychology: Research and Practice, 52*(6), 620–626. https://doi.org/10.1037/pro0000386

Funder, D.C., & Ozer, D.J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 56-168. doi:10.1177/2515245919847202

García-Pérez, M.A. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology, 8*, 100120. https://doi.org/10.1016/j.metip.2023.100120

Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*, 460-465. https://doi.org/10.1511/2014.111.460

Gergen, K. (1973). Social psychology as history. *Journal of Personality and Social Psychology*, *26*(2), 309-320. https://www.researchgate.net/publication/302871165_Social_Psychology_as_History

Gergen, K. (1994). Toward transformation in social knowledge (2nd ed.). Thousand Oaks, CA: Sage.

Goldacre, B., Drysdale, H., Dale, A. et al. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials* 20, 118. https://doi.org/10.1186/s13063-019-3173-2 Also see, https://www.compare-trials.org/

Goodwin, J.C. (2010). *Research in psychology: methods and design,* 6th Edition. Hoboken, NJ: Wiley.

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

Goulet-Pelletier, J.-C. & Cousineau, D. (2018) A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology*, *14*(4), 242-265. https://doi.org/10.20982/tqmp.14.4.p242

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

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130-143. http://dx.doi.org/10.1016/j.jmp.2015.09.001

Hays, W.L. (1994). *Statistics* (5th ed.). New York: Harcourt.

Hoaglin, D.C., Mosteller, F., & Tukey, J.W. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Hodges, J.L. & Lehmann, E.L. (1954). Testing the approximate validity of a statistical hypothesis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2), 261-268. https://doi.org/10.1111/j.2517-6161.1954.tb00169.x

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

Jackson, S.L. (2011). *Research methods and statistics: A critical thinking approach*, 4th edition. Belmont, CA: Wadsworth, Cengage Learning.

Julious, S.A. (2004) Tutorial in biostatistics - Sample sizes for clinical trials with normal data. *Statistics in Medicine, 23*(12). pp. 1921-1986. https://doi.org/10.1002/sim.1783

Jussim, L. (2017). Précis of *Social Perception and Social Reality*: Why accuracy dominates bias and self-fulfilling prophecy. *Behavioral and Brain Sciences*, *40*, 1-65. https://doi.org/10.1017/S0140525X1500062X

Kekecs, Z., Aczel, B., Palfi, B., Szaszi, B., Zrubka, M., Szecsi, P., Kovacs, M. & Bakos, B.E. (2020). Transparent Psi Project Stage 1 RR preregistration. OSF Registries. https://osf.io/a6ew3

Kekecs, Z., Palfi, B., Szaszi, B., Szecsi, P., Zrubka, M., Kovacs, M., Bakos, B. E., Cousineau, D., Tressoldi, P., Schmidt, K., Grassi, M., Evans, T. R., Yamada, Y., Miller, J. K., Liu, H., Yonemitsu, F., Dubrov, D., Röer, J. P., Becker, M., Schnepper, R., Ariga, A., Arriaga, P., Oliveira, R., Põldver, N., Kreegipuu, K., Hall, B., Wiechert, S., Verschuere, B., Girán, K., & Aczel, B. (2023). Raising the value of research studies in psychological science by increasing the credibility of research reports: The Transparent Psi Project. *Royal Society Open Science*, *10*: 191375. https://doi.org/10.1098/rsos.191375

Kennedy, J. E. (2023, March 4). Planning Falsifiable Confirmatory Research. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/pu2xy

Kennedy, J. E. (2023, April 19). Lessons and recommendations from a research audit for the Transparent Psi Project (TPP). PsyArXiv Preprints. https://doi.org/10.31234/osf.io/3wz6f

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

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572–584. https://doi.org/10.1177/1745691612463704

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. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Burlington, MA: Academic Press.

Kvarven, A., Strømland, E. & Johannesson, M. (2019). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behavior, 4*, 423–434. https://doi.org/10.1038/s41562-019-0787-z

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

Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician, 55*(3), 187-193. DOI: 10.1198/000313001317098149

Maier, M.A., Buechner, V.L., Dechamps, M.C., Pflitsch, M., Kurzrock, W., Tressoldi, P., Rabeyron, T., Cardeña, E., Marcusson-Clavertz, D., & Martsinkovskaja, T. (2020). A preregistered multi-lab replication of Maier et al. (2014), Exp. 4) testing retroactive avoidance. *PloS ONE, 15*(8): e0238373. https://doi.org/10.1371/journal.pone.0238373

Maxwell, S.E., Kelley, K., & Rausch, J.R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–63. http://dx.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,* 487-498. http://dx.doi.org/10.1037/a0039400

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

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

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

Meehl, P.E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108-141. https://doi.org/10.1207/s15327965pli0102_1

Meredith, M.P. & Heise, M.A. (1996). Comment on: "Bioequivalence trials, intersection–union tests and equivalence confidence sets" by Roger L. Berger and Jason C. Hsu. *Statistical Science, 11*(4), 304-306. https://projecteuclid.org/journals/statistical-science/volume-11/issue-4/Bioequivalence-trials-intersection-union-tests-and-equivalence-confidence-sets/10.1214/ss/1032280304.full

Meyners, M. (2012). Equivalence tests – A review. *Food Quality and Preference*, *26*(2), 231-245. https://doi.org/10.1016/j.foodqual.2012.05.003

Mitchell, G. (2012). Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, 7(2) 109 –117. https://doi.org/10.1177/1745691611432343

Mulder, J., & Wagenmakers, E-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments," *Journal of Mathematical Psychology, 72*, 1–5. http://doi.org/10.1016/j.jmp.2016.01.002

Murphy, K. R., Myors, B., & Wolach, A. (2014). Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (4th ed.). New York, NY: Routledge.

Nelson, L.D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534. https://doi.org/10.1146/annurev-psych-122216-011836

Nosek, B.A., Ebersole, C.R., DeHaven, A.C., & Mellor, D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences,* 115(11). https://doi.org/10.1073/pnas.1708274114

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

Pek, J. & Flora, D.B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23*(2), 208-225. http://dx.doi.org/10.1037/met0000126

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9,* 319–332. http://dx.doi.org/10.1177/1745691614528519

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pogrow, S. (2019). How effect size (practical significance) misleads clinical practice: The case for switching to practical benefit to assess applied research findings. *The American Statistician, 73*:sup1, 223-234. https://doi.org/10.1080/00031305.2018.1549101

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2023). Are small effects the indispensable foundation for a cumulative psychological science? A reply to Götz et al. (2022). *Perspectives on Psychological Science, 18*(2), 508–512. https://doi.org/10.1177/17456916221100420

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., … Schooler, J. (2020, September 10). High replicability of newly-discovered social-behavioral findings is achievable. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/n2a9x

Riesthuis, P., Mangiulli, I., Broers, N., and Otgaar, H. (2021). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology*. *36,* 203–215. https://doi.org/10.1002/acp.3911

Ritchie, S. (2020). *Science fictions: how fraud, bias, negligence, and hype undermine the search for truth*. New York: Henry Holt.

Rosnow, R.L. & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, *57*(3), 221-237. https://doi.org/10.1037/h0087427

Rosenthal, R., Rosnow, R.L., & Rubin, D.B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. New York: Cambridge University Press.

Rubin, M. (2021). When to adjust alpha during multiple testing: a consideration of disjunction, conjunction, and individual testing. *Synthese, 199,* 10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Schäfer, T. & Schwarz, M.A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*:813. https://doi.org/10.3389/fpsyg.2019.00813 doi: 10.3389/fpsyg.2019.00813

Schönbrodt, F.D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 1-15. http://dx.doi.org/10.3758/s13423-017-1230-y

Serlin, R.A., & Lapsley, D.K. (1993). Rational appraisal of psychological research and the good-enough principle. In: G. Keren & C. Lewis (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues* (pp. 199-228). Hilldale, NJ: Erlbaum.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U. (2015). The default Bayesian test is prejudiced against small effects. Data Colada Blog. Retrieved from http://datacolada.org/35

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stanley, T. D., Doucouliagos, H., & Ioannidis, J.P.A. (2022). Retrospective median power, false positive meta-analysis and large-scale replication. *Research Synthesis Methods*, *13*(1), 88-108. https://doi.org/10.1002/jrsm.1529

Strobe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science*, 7, 670–688. https://journals.sagepub.com/doi/pdf/10.1177/1745691612460687

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

Tong, C. (2019). Statistical inference enables bad science; Statistical thinking enables good science. *The American Statistician*, *73*, sup1, 246-261, https://doi.org/10.1080/00031305.2018.1518264

U.S. Food and Drug Administration (2010). *Guidance on the use of Bayesian statistics in medical device clinical trials*. Retrieved from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf

Vezzoli, M. & Zogmaister, C. (2023). An introductory guide for conducting psychological research with Big Data. *Psychological Methods*, *28*, 580-599. https://doi.org/10.1037/met0000513

Wagenmakers, E-J. (2015). A perfect storm: The record of a revolution. *The Inquisitive Mind*, Issue 25. http://www.in-mind.org/article/a-perfect-storm-the-record-of-a-revolution

Wagenmakers, E-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. (2012). An agenda for purely confirmatory research. *Perspectives in Psychological Science, 7*, 632–638. http://dx.doi.org/10.1177/1745691612463078

Wasserstein, R.L., Schirm, A.L., & Lazar, N.A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*, sup1, 1-19. https://doi.org/10.1080/00031305.2019.1583913

Watt, C.**,** & Kennedy, J. E (2015). Lessons from the first two years of operating a study registry. *Frontiers in Psychology, 7*, 173. https://doi.org/10.3389/fpsyg.2015.00173

Watt, C., & Kennedy, J. E. (2017). Options for prospective meta-analysis and introduction of registration-based prospective meta-analysis. *Frontiers in Psychology, 7*:2030. https://doi.org/10.3389/fpsyg.2016.02030

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p* hacking. *Frontiers in Psychology, 7*:1832. http://dx.doi.org/10.3389/fpsyg.2016.01832