How to Plan Falsifiable Confirmatory Research

James E. Kennedy and Caroline A. Watt

Version of June 14, 2018
Published on the internet in pdf and html at
https://jeksite.org/psi/falsifiable_research.pdf and https://jeksite.org/psi/falsifiable_research.htm

Abstract

Psychologists generally recognize falsifiable research as a basic goal of science. However, the methods for conducting falsifiable research with classical statistics and related methods for planning optimal Bayesian analyses have not yet been recognized and implemented by psychological researchers. The first step for falsifiable research is selection of a minimum effect size of interest such that a smaller effect would be too small to be of interest or would be evidence the hypothesis is false. If a minimum effect of interest is not specified explicitly, the effect size that just meets the criterion for acceptable evidence will implicitly function as a minimum effect of interest (e.g., the effect size that gives p = .05 or Bayes factor = 3). For confirmatory research, researchers should know what effect size is functioning as the minimum effect of interest. The second step is to determine the sample size that has power of at least .95 for the minimum effect of interest. Failure to obtain a significant result with power of .95 is evidence that the predicted effect specified in the power analysis is false for the conditions of the study. Such a failure can be considered as rejecting the alternative hypothesis at the .05 level using logic analogous to rejecting the null hypothesis. Evaluating the operating characteristics or power curve for a planned analysis reveals the effect sizes that can be reliably detected in a study and is needed for Bayesian methods as well as for classical methods. Specifying the effect sizes that can be reliably detected is as important as specifying the subject population. The third step is to publicly preregister the study with specific numerical inference criteria for evidence that the effect does not occur in addition to the usual criteria for evidence that the effect does occur. Studies with lower power may be conducted but the effect size with power of .95 is the *falsifiable effect size* for a study. Recent large studies have had adequate sample sizes for these methods. The relationships between these methods and meta-analyses, the "new statistics," and common practices for power analysis are discussed. Falsifiable research provides a conceptual framework for resolving many debates about methodology for confirmatory research.

Keywords: falsification, power analysis, confirmatory research, replication, minimum effect of interest, falsifiable hypothesis, operating characteristics, sample size, new statistics, registered replication reports

Testing falsifiable predictions is a basic goal of science and is generally accepted as a fundamental scientific principle in psychological research (Earp & Trafimow, 2015; Ferguson & Heene, 2012; Morey, Rouder, Verhagen, & Wagenmakers, 2014). A hypothesis or theory is scientifically weak and can be expected to be controversial if empirical tests that the hypothesis or theory is false are impossible or are not done. Statistically based research is biased if it can provide evidence that a hypothesis is true but cannot provide evidence that the hypothesis is false. As discussed below, this bias has been common is psychological research.

Falsifiable predictions are an appropriate goal for confirmatory research, but typically are not feasible or expected for the exploratory or discovery stage of research. We propose that *falsifiable predictions combined with public preregistration of all analysis decisions that could affect the primary study outcome can be considered defining characteristics of good confirmatory research* (Watt & Kennedy, 2015).

The recent "crisis of confidence" in psychological research revealed a lack of meaningful confirmatory research (Pashler & Wagenmakers, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) and implies a corresponding lack of falsifiable research. The practices that create false positive biases in research make research unfalsifiable. However, the challenges for implementing falsifiable research are deeper than the questionable research practices (John, Loewenstein, & Prelec, 2012) and researcher flexibility or degrees of freedom (Simmons, Nelson, & Simonsohn, 2011) that have been the focus of the crisis of confidence. The ongoing debates about the statistical methods and criteria for determining whether a replication study is successful reveal a fundamental lack of consensus about what constitutes confirmation and falsification (Anderson et al., 2016; Cumming, 2014; Gilbert, King, Pettigrew, & Wilson, 2016; Lakens, 2016; Maxwell, Lau, & Howard, 2015; Morey et al., 2015; Open Science Collaboration, 2015; Simonsohn, 2015a, 2016).

One overriding issue is that the statistical rationale for doing falsifiable research with classical statistics has generally been unknown or overlooked by psychological researchers. The common abuses of null hypothesis tests that have become widely recognized (Cumming, 2014) are a manifestation of the lack of attention to methods for falsifiable research. Proponents of Bayesian analysis have been the most outspoken advocates for falsifiable research (Dienes, 2014; Morey et al., 2015; Mulder & Wagenmakers, 2016).

The present paper discusses methods for conducting falsifiable research using classical statistics, and related methods that are needed for optimal application of Bayesian analysis for confirmatory research. The role of these methods in planning optimal confirmatory research has generally not yet been recognized by psychological researchers and underlies many of the ongoing debates about methodology for replication studies. After these methods are presented, we discuss how these methods relate to other types of data analysis, including meta-analysis, the "new statistics" as advocated by Cumming (2014), and Registered Replication Reports. We start with a discussion of the goals of falsifiable research and then discuss the three steps for planning falsifiable confirmatory research.

Realistic Goals for Falsifiable Research

Falsifiable research methods are applicable when scientists have theoretical ideas with empirically testable implications. These methods are not applicable when the goals of research are purely descriptive without a theoretical basis or prediction, as can occur for example with public opinion surveys.

We use the terms falsifiable prediction and falsifiable hypothesis interchangeably in this paper to mean that evidence can be obtained that the predicted effect does not occur for the conditions of the study. The qualifier *for the conditions of the study* indicates that the effect may not be absolutely false under all conditions. It is virtually always true in behavioral science that some aspects of the experimental conditions are different when a study is conducted by different researchers and/or at a different time. These differences could conceivably affect the study outcome.

Confirmatory research ultimately evaluates whether researchers have sufficient understanding to produce reliable demonstrations of an effect. If researchers cannot demonstrate reliable effects, the researchers do not have a convincing understanding of the effect and inferences about the effect are premature.

Also, providing *evidence* that a hypothesis is false does not imply that one study can provide compelling proof that a hypothesis is false. Confident inferences must be based on multiple confirmatory studies by different researchers (Earp & Trafimow, 2015). Each confirmatory study contributes to the overall evidence for a line of research and should to the extent possible provide strong unbiased evidence. We also focus in this paper on an effect or statistical hypothesis that is predicted to occur in a statistically based study and do not consider the question of whether the effect actually is evidence for the theoretical hypothesis or construct proposed by the researchers.

The pivotal question for a falsifiable prediction is: What inference or evidence can be drawn from a nonsignificant result? To the extent possible, the interpretation of a nonsignificant confirmatory study outcome should not be confounded by questions about the validity and reliability of the study methods. For example, if new measurement methods are being developed and evaluated as part of the study, the research is usually exploratory. Confirmatory research will typically use measurement methods that are established or have precedent.

Step 1. Select the Minimum Effect of Interest

The first step in planning falsifiable research is to select an effect size that is the minimum effect of interest such that a smaller effect would be evidence the effect does not occur or is too small to be of interest (Dienes, 2014). The minimum effect size of interest can be based on effects found in previous research or on a minimum effect of practical or theoretical interest. If researchers cannot identify a minimum effect size of interest, the research remains at an exploratory stage and has not developed to the point of a falsifiable theory and associated predictions. The rationale for this point is explained in the section below on power analysis.

Effects Found in Previous Research

If the primary goal of a study is to confirm the effects found in previous research, the minimum effect of interest will typically be based on the confidence interval from the previous research. The 95% or 99% confidence interval, or possibly an even more extreme value, may be an appropriate minimum effect of interest. One-sided confidence intervals may be applicable.

However, the use of effects from previous research can be severely limited by wide confidence intervals from small sample sizes in previous studies and by the likelihood of bias from methodological flexibility and questionable research practices. The effects in previous studies may not be useful predictors of expected effects for the first confirmatory studies in a line of research.

Minimum Effect of Practical or Theoretical Interest

The optimal strategy for planning confirmatory research is to identify a minimum effect size of practical or theoretical interest. This effect size is based on overall knowledge of the subject matter and the recognition that previous findings may reflect a real effect that was artificially enhanced by methodological bias. The fact that previously reported larger effects do not replicate does not mean an effect is completely invalid.

The important scientific question is what minimum effect size would be meaningful? Psychological researchers struggle with this question (Maxwell, Kelley, & Rausch, 2008) and often appear to treat statistical significance as establishing that an effect is meaningful. However, we suggest that researchers should consider statistics as a tool to evaluate whether the data in a study are consistent with a pre-specified meaningful effect size and not as defining which effect sizes are meaningful. This distinction underlies many of the abuses of null hypothesis testing.

Cohen (1988) proposed conventions for “small effects” that would typically be the smallest effect size that a behavioral researcher might want to investigate and could reasonably distinguish from zero. He specified these small effects as an effect size equivalent to a correlation coefficient of .10 (accounting for 1% of the variance) or the equivalent Cohen’s *d* of .20. This effect size may be an appropriate minimum effect of practical or theoretical interest for many areas of behavioral research.

However, Cohen also pointed out that the effect size of interest will be different for different topics of research. Effect sizes smaller than a correlation of .10 may be relevant for effects that could directly affect human health and welfare for millions of people; however, that type of research is rare for academic behavioral research. Similarly, larger effect sizes may be appropriate for some areas of research. For example, an effect equivalent to a correlation of .14 or .20 (accounting for 2% and 4% of the variance) and the corresponding Cohen’s *d*s of .28 and .41 may be an appropriate minimum effect of practical or theoretical interest when more conspicuous effects are predicted.

Minimum Effect of Interest in Bayesian Analysis

Minimum effects of interest are applicable for Bayesian analysis as well as for classical analyses. Kruschke's (2015) strategy of identifying a "Region of Practical Equivalence (ROPE)" for Bayesian analysis is the same concept as a minimum effect of practical or theoretical interest.

Dienes (2014) argued that a benefit of using the Bayes factor for data analysis is that falsifiable tests can be conducted without specifying a minimum effect of interest. However, we point out that a Bayes factor analysis translates observed effect sizes to odds. The magnitude of odds that is considered to be acceptable evidence (often 3) is associated with a corresponding effect size that in practice becomes the minimum effect of interest. The fact that researchers typically do not know what effect size is functioning as the minimum effect of interest does not mean that a minimum effect is not involved.

Our perspective is that effect size is the outcome of practical interest for a study and the medium of exchange for research. The best strategy for confirmatory research is to identify a minimum effect of interest and then design a study that evaluates whether the data are consistent with that effect. As noted above, a common unfortunate practice in past psychological research has been to largely ignore effect size and assume that the *p* value indicates whether an observed effect is meaningful. Shifting from *p* values to Bayesian odds with a similar disregard for effect size can be expected to have continued unfortunate consequences. We recommend that effect size be in the forefront when planning and interpreting studies. For a Bayes factor analysis, methods described below can be used to identify the effect sizes that can be reliably detected in a planned study and thus the effects that are actually investigated with the study.

Step 2. Evaluate the Operating Characteristics or Power Curve

Power analysis is the mechanism for implementing falsifiable predictions when using classical statistics. Power analysis determines the probability that a planned statistical analysis will correctly detect an effect if the true effect size is a certain value. The effect size specified in a power analysis is a prediction about the study outcome. As discussed below, failure to obtain the expected outcome in a well-powered study is evidence that the prediction was wrong. Without a power analysis, a nonsignificant outcome cannot provide evidence that a hypothesis is false because the outcome could be due to low power. Thus, a study with low power is biased research that can provide evidence that a hypothesis is true but cannot provide evidence that the hypothesis is false. Most discussions of power analysis in psychology have focused on setting sample size and have not discussed power analysis in context of falsifiable predictions.

The typical null hypothesis tests used in psychological research in recent 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 consistent with exploratory research. They are not falsifiable in principle because any finite amount of data could have inadequate

power to detect the extremely small effects that are consistent with the hypotheses. In the absence of a minimum effect of interest these hypotheses are impervious to falsification.

On the other hand, many researchers apparently do not understand the role of power analysis and incorrectly interpret a nonsignificant outcome with low or unknown power as evidence that an effect does not occur (Anderson & Maxwell, 2016; Cohen, 1988, p. 16; Dienes, 2014; Finch, Cumming, & Thomason, 2001). The lack of attention to statistical power resulted in studies with very low power. Bakker, van Dijik, and Wicherts, (2012) estimated that psychological studies have had an average power of .35 and Button et al. (2013) estimated that neuroscience studies have had a median power of .21.

Power analysis reveals the effect sizes that can be reliably detected in a study—and thus the effect sizes that are actually evaluated by the study and implicitly predicted by the researchers. For confirmatory research, researchers should know what effect sizes are actually being investigated and should interpret the results accordingly.

Operating Characteristics or Power Curve

The operating characteristics or power curve for a planned statistical analysis provides clear information about which effect sizes can be reliably detected and which cannot. The operating characteristics are a graph or table that gives the probability of detecting the effect if the true effect size is a certain value. These probabilities are determined for the range of effects that could possibly occur in the study. Operating characteristics may be derived mathematically or from simulations of the planned analyses, including for sequential analyses.

The operating characteristics quantitatively evaluate the expected rates of inferential errors or statistical validity of a planned statistical inference and should be considered a component of standard confirmatory research methodology. Specifying the effect sizes that can be reliably detected in a planned study is as important as specifying the subject population. Operating characteristics are applicable whether an inference is based on a statistical hypothesis test or on an evaluation of whether the confidence interval contains or excludes a particular value.

Operating characteristics reveal the effect sizes that can be reliably detected with Bayesian analysis as well as with classical analyses. As noted above, Bayesian hypothesis tests, like classical hypothesis tests, may inspire an overemphasis on the probability of an outcome and a lack of attention to effect size. In addition, for Bayesian methods prior probability distributions can have practical implications that are not easy to understand and can introduce substantial biases that make a hypothesis test have unintended and unwanted properties (Gu, Hoijtink, & Mulder, 2016; Kennedy, 2015; Kruschke, 2015; Simonsohn, 2015b). The operating characteristics make the practical implications of prior probability distributions clear and reveal potential biases. Operating characteristics are expected for Bayesian analyses used in regulated medical research in the U.S. (U.S. Food and Drug Administration, 2010), and that is the most well developed discussion of Bayesian analysis for confirmatory research that we have found.

Power analyses (Kruschke, 2015) and similar *design analyses* (Schönbrodt & Wagenmakers, 2016) have been described for Bayesian analyses in psychological research. These discussions focus on estimating sample size and inferential error rates for study planning. However, they have not addressed the full operating characteristics described above that are useful for identifying the effect sizes that can be reliably detected in a study. The study-planning analyses described by Krushchke (2015) and by Schönbrodt and Wagenmakers (2016) are based on simulations that sample from a prior probability distribution of possible effect sizes, whereas the simulations to evaluate the effect sizes that can be reliably detected sample from models of individual effect sizes that cover the range of effect sizes of interest. The latter approach more clearly displays the relationship between effect size and evidence, and provides much more useful information for evaluating the planned analysis. The research plan developed by Kekecs and Aczel (2018) used both methods when planning a sequential Bayesian analysis and is a good model for effective statistical planning.

Classical Evidence that the Alternative Hypothesis is False

The effect size with a power of .95 for the alternative hypothesis can be rejected or falsified using the same .05 criterion applied for the null hypothesis. Cohen (1988, pp. 16-17) briefly described this statistical principle during an earlier methodological era when exploratory and confirmatory research were not clearly distinguished. He presented it as a theoretical point rather than as a practical method for conducting falsifiable research.

Figure 1 displays the distribution for the null hypothesis with the significance level or alpha set to .05 and the distribution for the alternative hypothesis with a power of .95. A nonsignificant outcome can be viewed as rejecting the alternative hypothesis at the .05 level in the same way that a significant outcome is viewed as rejecting the null hypothesis. If consistency of terminology is wanted, such analyses with high power are equivalent to switching the null and alternative hypotheses, and rejecting the new null hypothesis that the effect size is the size specified in the original power analysis.

Power of .95 is the clearest example, but the same principle is applicable for lower power such as .90 and for higher power. Note that rejecting the specified alternative hypothesis is also rejecting all possible alternative hypotheses to the right of the specified minimum alternative hypothesis because those hypotheses would have even greater power.

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.

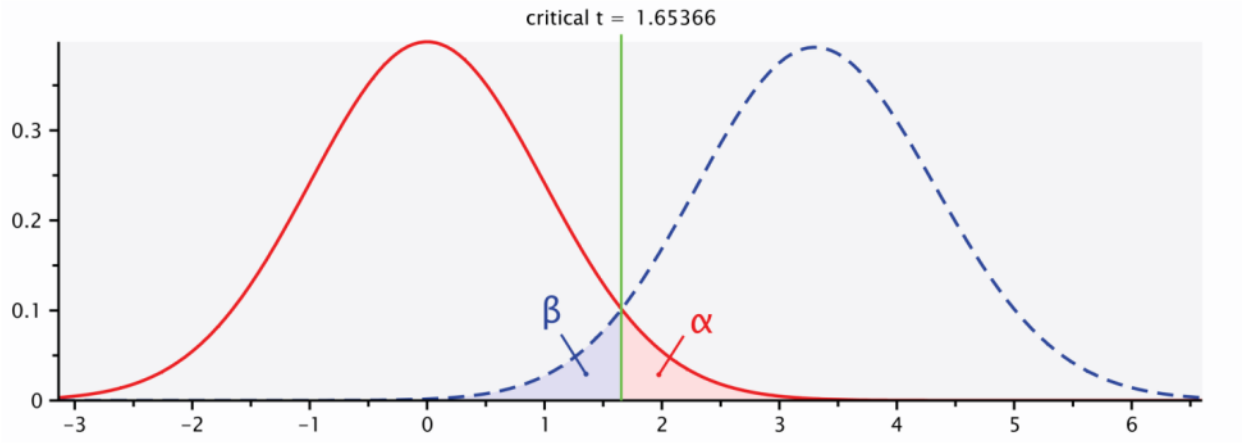

Figure 1. Distributions for the null hypothesis (left, solid line) and alternative hypothesis (right, dashed line) with α = .05 and β = .05 for power = 1 – β = .95. An outcome to the right of the vertical line is significant and to the left is nonsignificant. The plot is from the G*Power program (Faul, Erdfelder, Lang, & Buchner, 2014).

Bayesian Evidence that the Alternative Hypothesis is False

Bayesian methods can also provide evidence that an experimental hypothesis is false. The most common method is the Bayes factor that quantifies and compares the extent to which the observed data would be expected with the alternative model and 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 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 (Schönbrodt & Wagenmakers, 2016). As mentioned above, a Bayes factor analysis does not require explicit specification of a minimum effect of interest; however, the operating characteristics reveal the effect sizes that can be reliably detected in a planned study and thus the effects that are treated as being of interest.

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 classical 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. Additional cautions with Bayes factor analyses are discussed in the section below on evidence that the null hypothesis is true.

Differences from Common Practices for Power Analysis

Psychological researchers have frequently assumed that a power analysis must be based on the true effect size for an effect—which has resulted in technical discussions about the difficulty in estimating the true effect size and the implications for power analysis (Maxwell et al, 2015; McShane & Böckenholt, 2016). One common practice has been to base a power analysis on the mean effect size from previous studies. This is not recommended because it does not realistically consider the uncertainty of the estimates (Perugini, Gallucci, & Costantini, 2014; Yuan & Maxwell, 2005). It also does not consider the likelihood of biased estimates due to questionable research practices and researcher flexibility. The assumed need to specify the unknown true effect size for power analysis has been one of the greatest obstacles to the implementation of power analysis (Maxwell et al., 2008).

The approach discussed here of identifying a minimum effect size of interest that the researchers want to be able to detect reliably is more practical. Effect sizes larger than this minimum will have greater power and thus will have adequate power if this minimum has adequate power. This strategy addresses the uncertainties and inexact hypotheses of research and has long been recognized (Hays, 1973, pp. 362-365). Rather than focusing on a precarious point estimate for the power of a study, this strategy assures that the relevant range of possible values of effect size has adequate power.

Power of .80 has often been considered adequate for setting sample size; however, a power closer to .95 or greater is needed for falsifiable research. Cohen (1965) originally recommended a power of .80 in the 1960s when typical data collection and many analyses were done manually. 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 the alpha or significance level of 05. However, a .20 probability of failing to detect a true effect does not allow confident inferences that an effect is false for a study. With current technology, data collection and analyses are typically much more efficient than in the 1960s and larger studies are now feasible. For confirmatory research as discussed here we recommend that false positive and false negative results generally be considered about equally undesirable, and that inference criteria be set accordingly.

A confirmatory analysis will usually be one-sided. If researchers cannot predict whether an effect will be positive or negative, the research is usually still at an exploratory stage rather than confirmatory research. Similarly, researchers using general linear models or ANOVAs will normally use more powerful directional contrasts for confirmatory hypotheses rather than the intrinsically two-sided and more exploratory ANOVA main effects and interactions (Rosenthal, Rosnow, & Rubin, 2000).

The Dilemma of Heterogeneity

McShane and Böckenholt (2014) pointed out that heterogeneity or between-study variation in effect sizes due to unrecognized factors can make power estimates overly optimistic, particularly for small effects. They also noted that heterogeneity can occur even when every effort is made to duplicate the methods used in previous research. The discussion of heterogeneity is basically a technical presentation of the conceptual point that confirmatory research evaluates whether researchers have sufficient understanding to demonstrate reliable effects. Heterogeneity can be considered a measure of the extent to which researchers do not understand the phenomena they are investigating. If the results are inconsistent, the researchers do not have adequate understanding, and inferences about the effects are premature. As McShane and Böckenholt (2014) recommended, multicenter studies are the best strategy to establish the generality of findings and to identify factors that may influence the effects.

Inferences about causality must be made with caution in the presence of substantial heterogeneity. When heterogeneous effect sizes are found, "it is unclear whether the various research findings represent a common underlying phenomenon" (Wood & Eagly, 2009, p. 459). Substantial heterogeneity is an indication that important variables are not yet understood and controlled—and therefore are potential confounding factors. Large sample sizes do not resolve the potential confounding if important causal variables are not yet understood and controlled.

Step 3. Preregister the Statistical Methods and Inference Criteria

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 from the data. For Bayesian methods, the selected prior probability distributions should also be included in the preregistration.

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 a *p* value, a Bayesian odds, or a confidence interval or Bayesian credible interval that is used to make inferences. In all cases, the observed confidence intervals or credible intervals should be carefully considered relative to the minimum effect size of interest to obtain a practical understanding of the observed effect.

Similar methods can be used for the criteria for evidence that the alternative hypothesis is false. For a classical hypothesis test with a power of .95 or larger, a nonsignificant outcome is

evidence that the effect size specified in the power analysis is false. Alternatively, a confidence interval that does not contain the minimum effect of interest or a larger value is evidence that the minimum effect of interest is not true. One-sided confidence intervals may be appropriate for these evaluations. Corresponding inferences can be made with Bayesian hypothesis tests and credible intervals. For classical hypothesis tests a significance level more extreme than .05 may be useful in some cases. The optimal practice in such cases would typically be to set power to one minus the significance level (e.g., for significance level of .01, power would be set to 1 - .01 = .99). However, the significance level and power need not precisely correspond like this if the operating characteristics for the analysis have useful properties.

Inconclusive outcomes are unlikely with carefully planned, well-powered confirmatory studies, but 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, and that situation merits special consideration.

If Power of .95 is not Feasible

In some situations researchers may not have adequate resources to conduct a confirmatory study with a power of .95 for the minimum effect of interest. We suggest that in such cases the researchers conduct studies with the available resources but openly acknowledge the limitations of the study. Studies with lower power increase the probability of producing inconclusive results, but the study outcome could possibly meet the criteria for drawing inferences about the experimental hypothesis. Presenting the operating characteristics that identify the power for the different possible effect sizes is particularly useful in these cases.

The preregistration in this situation should note the power for the minimum effect of interest and the effect size that has power of .95. The effect size with power of .95 can be considered to be the *falsifiable effect size* for the study. This is a useful descriptor for any study. The gap between the falsifiable effect size and the minimum effect of interest is a region of compromised power. If the true effect lies within this region, the study has a substantial chance of producing inconclusive results.

For studies with power less than .95, confidence intervals will normally be needed to interpret a nonsignificant result. If the confidence interval contains both the minimum effect of interest and the value for the null hypothesis, the outcome is inconclusive. However, if the confidence interval does not contain the minimum effect of interest or a larger value, the outcome is evidence that the minimum effect of interest is not true (which, as discussed below, is not the same a concluding that the null hypothesis is true).

Studies with power not far below .95 have relatively little risk of inconclusive results. For example, a study with power of .90 may be generally acceptable for confirmatory research. Because the power is less than .95, the preregistered inference criteria would specify the evaluation of confidence intervals to interpret a nonsignificant outcome.

Relation to other Statistical Methods

Meta-Analysis

The principles of preregistered confirmatory research and falsifiable hypotheses apply to meta-analyses as well as to individual studies. Although meta-analysis is often viewed as the definitive form of research synthesis, the flexibility and retrospective nature of the methodological decisions in typical retrospective meta-analyses are characteristics of exploratory research rather than confirmatory research (Watt & Kennedy, 2017). Retrospective meta-analysis is a form of post hoc analysis in which the methodological decisions about the criteria for including studies, statistical methods, and moderating variables are made after the outcomes are known for the studies that may be included. The flexibility in making these decisions introduces substantial potential for bias.

One of the prominent manifestations of the exploratory nature of retrospective meta-analyses is that they have rarely been effective at resolving scientific debates (Ferguson, 2014; Ferguson & Heene, 2012; van Elk et al., 2015). After noting several debates in psychology that meta-analyses failed to resolve, Ferguson (2014, p. 1) commented that these were "a small subset of examples" and that "It may be that meta-analyses are not capable of answering big questions due to issues of methodological flexibility."

On the other hand, *prospective meta-analyses* can be conducted as confirmatory research with preregistration and falsifiable predictions established for the meta-analysis before data collection begins for the included studies. We have outlined options and methodological recommendations for prospective meta-analysis elsewhere (Watt & Kennedy, 2017). Prospective meta-analysis may be most useful in resolving scientific debates and particularly in establishing evidence that an effect does not occur.

The Registered Replication Reports (RRR) originally developed for the journal *Perspectives on Psychological Science* and now transferred to *Advances in Methods and Practices in Psychological Science* are prominent examples of prospective meta-analysis in psychological research (Association for Psychological Science, n.d.; Simons, Holcombe, & Spellman, 2014). The detailed protocol for conducting the studies is developed before data collection begins for the included studies, and thus, many potential biases from retrospective methodological decisions are eliminated. The recommended data analysis methods for RRR are based on the "new statistics" advocated by Cumming (2014) and these methods are discussed below.

Evidence that the Null Hypothesis is True

The research strategies discussed in this paper focus on evidence that the predicted alternative hypothesis is true or false and do not address evidence that the null hypothesis is true. Evidence that the alternative hypothesis is false is usually not strong evidence that the null hypothesis is true because the null and alternative hypotheses could both be false. For example, the true effect size could be close to zero, but not zero. Or, for a one-sided test the true effect

could be in the opposite direction. In order to have unambiguous evidence that the null hypothesis is true, a range of values needs to be identified that is considered indistinguishable under the null hypothesis. The confidence interval for the effect size estimate must be entirely within this range. The upper boundary for this range may or may not be the minimum effect of interest for evaluating the alternative hypothesis. This type of analysis is *equivalence testing* and generally requires larger sample sizes than evaluating an alternative hypothesis. Maxwell et al. (2015) and Lakens (2017) provide overviews of these methods.

If a standard equivalence test is adapted by (a) making the test one-sided rather than two-sided, (b) setting the equivalence boundary equal to the minimum effect of interest as described in this paper, and (c) setting the sample size to have power of .95 if the true effect size is zero, the test will give the same inferences about whether the minimum effect of interest is not true as the methods recommended in this paper.

We do not discuss equivalence testing further because we think that most confirmatory research will (or should) focus on evaluating the predicted alternative hypothesis rather than confirming the null hypothesis. With the falsifiable research strategy discussed here, the null hypothesis is not of interest other than for developing a test of the alternative hypothesis.

This point is worth reiterating because it is so different from how most researchers have been thinking about hypothesis tests. Falsifiable research focuses on evidence that the predicted alternative hypothesis is true or false and is not an artificial binary choice between the null and alternative hypotheses. Evaluating whether the null or some other model is applicable if the predicted alternative is false would be exploratory analyses beyond the primary confirmatory analysis.

The possibility that neither the alternative nor null models accurately capture the effect is also applicable to Bayesian analyses that compare null and alternative models. In general, the fact that the data fit the null model better than an alternative model could be due to the selected alternative being far from optimal rather than the null model being precisely true. As shown in the evaluation of inferential errors in the study plan by Kekecs and Aczel (2018) (section on Operational Characteristics), a careful evaluation of operating characteristics will reveal the effect sizes for which a planned analysis tends to incorrectly support the null model when it is not true. The safest inferences focus on the validity of the predicted alternative model and associated minimum effect of interest without attempting to infer that the null model is precisely true.

With classical statistics, inferences based on hypothesis tests and inferences based on confidence intervals are developed from the same underlying statistical theory and give the same outcomes (Hays, 1973; Kutner, Nachtsheim, Neter, & Li, 2005). The choice between the null and alternative models in a classical hypothesis test developed with power analysis can be viewed as a convenient method for identifying efficient test conditions that provide useful confidence intervals to evaluate predictions.

However, Bayes factor hypothesis tests do not have a similar equivalence relationship with credible intervals (the Bayesian counterpart to confidence intervals). Bayes factors can give results that are different from reasonable inferences from estimation methods based on credible intervals (Kruschke, 2015). Thus, additional effort is required to establish the relationship between effect size and Bayes factor odds when planning a study. Inferences based on Bayesian estimation methods with an explicit ROPE are more direct and may be more intuitive.

The "New Statistics"

The "new statistics" advocated by Cumming (2014) focus on estimating effect sizes and confidence intervals and explicitly avoid "dichotomous thinking" and associated hypothesis tests. Cumming's recommended solution to the past abuses of hypothesis tests was to abandon hypothesis tests to the maximum extent possible. As noted above, we generally agree with the central role of effect sizes as emphasized in the new statistics, and that framing an inference as a binary choice between two models can be more of a distraction than benefit if the implications for effect size are not carefully considered.

However, a general avoidance of dichotomous thinking and hypothesis tests appears to imply an avoidance of falsifiable predictions (Kennedy, 2016a), which would be a significant step backward for scientific research. In response to Cumming's paper, Morey et al., (2014) commented:

> For psychological science to be a healthy science, both estimation and hypothesis testing are needed. Estimation is necessary in pretheoretical work before clear predictions can be made, and is also necessary in posttheoretical work for theory revision. But hypothesis testing, not estimation, is necessary for testing the quantitative predictions of theories. (p. 1290)

We believe that the falsifiable research methods proposed here integrate the strengths of the new statistics with the strengths of hypothesis tests. Rather than repeat and elaborate the points of academic debate, it may be more useful to examine how the new statistics have played out in practice.

The Registered Replication Reports noted above in the discussion of meta-analysis are intended to apply the principles of the new statistics. Simons et al., (2014) stated that "The data analysis concentrates on estimation of effect sizes rather than on significance testing" (p. 553-554). The RRR webpage further states that "The emphasis on estimating effect sizes rather than on the dichotomous characterization of a replication attempt as a success or failure based on statistical significance could lead to greater awareness of the shortcomings of traditional null-hypothesis significance testing" (Association for Psychological Science, n.d.).

As of November, 2017, six RRRs have been published and all six described the results in ways that clearly stated or implied that the RRR either successfully confirmed (one case) or failed to replicate (five cases) the previous findings. Alogna et al., (2014) concluded that their RRR findings "provide clear evidence for" the effect (p. 571). Eerland et al., (2016) concluded

that “This RRR did not find” the effect (p. 158) and the results “are not consistent with the original result” (p. 166). Hagger et al., (2016) reported that “The results are consistent with a null effect” (p. 556) and “provide evidence that, if there is any effect, it is close to zero” (p. 558). Cheung et al., (2016) reported that their RRR results “are not consistent with the original result” (p. 761) and “This RRR did not find a causal effect” (p. 761); however, the interpretation of the outcome was confounded because a prerequisite manipulation failed to produce the expected effect. Wagenmakers et al., (2016) reported that “Overall, the results were inconsistent with the original result” (p.924) and “This RRR did not replicate the [previous] result and failed to do so in a statistically compelling fashion” (p. 924). Bouwmeester et al., (2017) reported that the RRR results “are consistent with the presence of selection biases and the absence of a causal effect” (p. 527) and “Overall, the results of the primary analysis in this RRR showed essentially no difference … the point estimate was opposite that predicted by the hypothesis and was close to zero” (p. 537).

The experience with RRRs suggests that inferences about whether a confirmatory study successfully confirmed or failed to replicate the previous findings are virtually inevitable. Our perspective is that such dichotomous inferences are the primary purpose of conducting well-designed confirmatory research—and that dichotomous thinking that distinguishes between what is true and what is not true is the goal of empirical science in general.

Given that inferences about the success of a confirmatory study can be expected, one major effect of the estimation approach in the new statistics has been that the inferences are less structured than for hypothesis tests and the inference criteria appear to have been developed after looking at the data rather than preplanned and preregistered. Leaving the inference criteria to be decided after the results are known is significant researcher flexibility (degrees of freedom) that is typical for exploratory research, but is not consistent with the emerging standards for confirmatory research (Wagenmakers et al., 2012; Wicherts et al., 2016).

Attention to efficiency of research is another difference between the implementation of the new statistics and the falsifiable research methods discussed here. The overall numbers of subjects in the six RRRs were 3596 (Bouwmeester et al., 2017), 2284 (Cheung et al., 2016), 2141 (Hagger et al., 2016), 2055 (Alogna et al., 2014), 1894 (Wagenmakers et al., 2016), and 927 (Eerland et al., 2016). The sample sizes for five of the RRRs were substantially larger than the approximately 1100 subjects noted above as needed for .95 power for a small effect size. The RRRs demonstrate that studies with sufficient sample size for falsifiable research are feasible in the current research environment.

Efficiency of research is one of the major goals and selling points for power analysis and hypothesis tests (Cohen, 1988; Hays, 1973; Schönbrodt & Wagenmakers, 2016). Properly conducted power analyses and hypothesis tests are intended to avoid wasting research resources by doing studies that are too small to make the points the researchers want to make and also avoid wasting resources by doing studies that are much larger than needed to make those points. The falsifiable research methods discussed here, particularly the evaluation of operating

characteristics, implement this efficiency without losing sight of the primary importance of effect sizes.

If resources are available for studies with about 2000 or more subjects as occurred with five of the RRRs, psychological researchers usually are reasonably safe in not conducting a power analysis. However, if resources are limited and efficient allocation of research effort is a consideration, power analysis should have a central role in the design of confirmatory studies. We believe that consideration of efficiency is generally a preferable strategy for psychological research. In addition, given the current limited understanding of potential biases with Bayesian analyses, we recommend that the evaluation of operating characteristics is done for any confirmatory inferences based on Bayesian methods.

Final Thoughts

The methodological biases that have received much attention in recent years fit within the conceptual framework of failing to conduct falsifiable research. We believe that psychological research will remain methodologically compromised until researchers implement confirmatory research with public preregistration and falsifiable predictions as described in this article.

Our observation is that the recent efforts to remedy the methodological weaknesses in psychology fall short of the research standards that are needed. The ready acceptance by many psychologists of the “new statistics” and the associated retreat from hypothesis tests and pre-specified inference criteria represent continued embrace of researcher flexibility (degrees of freedom) and under-emphasis of falsifiable confirmatory research. Similarly, recent articles discussing power analysis have predominantly focused on obtaining evidence that an effect occurs and have not considered study designs and inference criteria for evidence that a hypothesis is false. Implementation of falsifiable research requires significant changes from current common methodological practices.

The need for these methods is most obvious in controversial areas of research such as parapsychology (Watt & Kennedy, 2015; Wagenmakers et al., 2012). However, the controversies about parapsychology forcefully reveal methodological issues that must be recognized and addressed for psychological research in general (Wagenmakers et al., 2012).

In addition to the topics discussed in this article, psychological researchers generally have not yet recognized certain other methodological issues that will eventually have to be addressed in confirmatory research. These issues include the need to handle incomplete data, dropouts, and other protocol violations more carefully and more conservatively, the need for formal documented software validation, and the need to implement procedures that make experimenter fraud difficult (Kennedy, 2016b). As discussed in Kennedy (2016b), methods for addressing these issues have been developed in regulated medical research and can be adapted to psychological research.

References

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S., Birt, A. R., ... Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9,* 556–578. doi: 10.1177/1745691614545653

Anderson, C.J., Bahník, S., Barnett-Cowan, M., Bosco, F.A., Chandler, J., Chartier, C.R. … Zuni, K. (2016). Response to Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037. doi: 10.1126/science.aad9163

Anderson, S. F. & Maxwell, S.E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21,* 1–12. doi: org/10.1037/met0000051

Association for Psychological Science. (n.d.). Registered Replication Reports. Retrieved from https://www.psychologicalscience.org/publications/replication.

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

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

Button , K. S., Ioannidis, J. P. A, Mokrysz1, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376. doi:10.1038/nrn3475

Cheung, I., Campbell, L., LeBel, E., . . . Yong, J. C. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11,* 750–764. doi: 10.1177/1745691616664694

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.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25,* 7-29. Retrieved from doi: 10.1177/0956797613504966

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5,* 781. doi: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. doi:10.3389/fpsyg.2015.00621

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., . . . Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11,* 158–171. doi: 10.1177/1745691615605826

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2014). G*Power. [Software]. Retrieved from http://www.gpower.hhu.de/

Ferguson, C. J. (2014). Comment: Why meta-analyses rarely resolve ideological debates. *Emotion Review, 6*(3). doi: 10.1177/1754073914523046

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives in Psychological Science, 7*, 555-561. doi: 10.1177/1745691612459059

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210. doi: 10.1177/00131640121971167

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

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

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573. doi: 10.1177/1745691616652873

Hays, W.L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart, and Winston.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science, 23*, 524–532. doi:10.1177/0956797611430953

Kekecs, Z. & Aczel, B. (2018). Transparent Psi Project - Research Plan. Available at https://osf.io/d7sva/.

Kennedy, J. E. (2015). Beware of inferential errors and low power with Bayesian analyses: Power analysis is needed for confirmatory research. *Journal of Parapsychology, 78*, 170-182. Available at http://jeksite.org/psi/jp15.pdf .

Kennedy, J. E. (2016a). Critique of Cumming's "New Statistics" for psychological research: A perspective from outside psychology. Available at http://jeksite.org/psi/critique_new_stat.pdf.

Kennedy, J. E. (2016b). Is the methodological revolution in psychology over, or just beginning? *Journal of Parapsychology, 80*, 156-168. Available at http://jeksite.org/psi/methods_predictions.pdf

Kutner, M.H., Nachtsheim, C.J., Neter, J, & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York: McGraw Hill.

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

Lakens, D. (2016). The statistical conclusions in Gilbert et al. (2016) are completely invalid. The 20% Statistician Blog. Retrieved from http://daniellakens.blogspot.co.uk/2016/03/the-statistical-conclusions-in-gilbert.html

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8,* 355-362. doi: 10.1177/1948550617697177

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. doi: 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. doi: 10.1037/a0039400

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. doi: 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. doi: 10.1037/met0000036

Morey, R.D., Rouder, J.N., Verhagen, J., & Wagenmakers, E-J. (2015). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science, 25,* 1289–1290. doi: 10.1177/0956797614525969

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. doi: 10.1016/j.jmp.2016.01.002

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

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

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

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review,* 1-15. doi: 10.3758/s13423-017-1230-y

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

Simons, D.J., Holcombe, A.O., & Spellman, B.A. (2014). An introduction to registered replication reports at *Perspectives on Psychological Science*. *Perspective on Psychological Science, 9*, 552-555. doi: 10.1177/1745691614543974

Simonsohn, U. (2015a). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559–569. Retrieved from http://datacolada.org/wp content/uploads/2016/03/26-Pscyh-Science-Small-Telescopes-Evaluating-replication-results.pdf

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

Simonsohn, U. (2016). Evaluating replications: 40% full ≠ 60% empty. Data Colada Blog. Retrieved from http://datacolada.org/47

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

van Elk, M., Matzke, D., Gronau, Q. F., Guan, M., Vandekerckhove, J., & Wagenmakers, E-J. (2015). Meta-analyses are no substitute for registered replications: A skeptical perspective on religious priming. *Frontiers in Psychology, 6*:1365. doi: 10.3389/fpsyg.2015.01365

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., … Zwaan, R. A. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*, 917–928. doi: 10.1177/1745691616674458

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. doi: 10.1177/1745691612463078

Watt, C., & Kennedy, J. E. (2015). Lessons from the first two years of operating a study registry. *Frontiers in Psychology*, 7, 173. doi: 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. doi: 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. doi:10.3389/fpsyg.2016.01832

Wood, W., & Eagly, A. H. (2009). Advantages of certainty and uncertainty. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.) (pp. 455-472). New York: Sage.

Yuan, K-H., & Maxwell, S.E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30,* 141–167. doi: 10.3102/10769986030002141

About the Status of this Paper

The ideas in this paper were submitted in different forms to two psychological journals and both rejected the paper. In one case, the dominant reviewer was a strong advocate for the "new statistics" and wanted a major rewrite of the paper that included being more favorable to the new statistics and not pointing out that avoidance of dichotomous thinking implies avoidance of falsifiable predictions. We declined to make those changes. In the other case, the editor (who was also a proponent of the new statistics) said that power analysis and related topics have been extensively discussed in the existing literature and this paper was not sufficiently novel for publication in their journal. The manuscript was not sent out for review. Our observation is that the ongoing debates about statistical methods for replication studies and about the value of retrospective meta-analysis show that methodological opinions are heading in many different directions and not converging to a consensus. The debates between advocates for the new statistics and advocates for hypothesis tests are a clear example. We believe that falsifiable research provides a conceptual framework for resolving these debates and implementing optimal research methods. However, we have not found existing articles that provide useful discussions of the rationale and practical methods for implementing falsifiable research. We believe that the present paper provides valuable guidelines that are needed in psychological research—whether or not the various ideas in the paper are considered novel. Given our experience with these gatekeepers for psychology, we decided the best approach is to publish the paper as open access on the internet—which also gives us full control of the copyright and distribution of the paper.

Author Notes

James E. Kennedy is retired from a professional career that involved diverse experience in research and data analysis in academic, nonprofit, government, and industry settings. His views on research methodology have been particularly influenced by working in regulated medical research. Email jek@jeksite.org

Caroline A. Watt is Professor of Psychology in the School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK. Email Caroline.Watt@ed.ac.uk

Copyright Notice

Revision History

January 24, 2018 – Article initially posted on the internet.

May 21, 2018 – Added the Kekecs and Aczel (2018) reference on pages 7 and 13, and the sentence on page 10 about heterogeneity being a measure of the extent to which researchers do not understand the phenomena they are investigating.

May 22, 2018 – Added the sentence on page 10 that "In all cases, the observed confidence intervals or credible intervals should be carefully considered …"

June14, 2018 – Added sentence on page 13 about the conditions for an equivalence test to give same inferences as the methods recommended in this paper.

Other Methodology Articles