Journal of Behavioral
Public Administration
Vol 8, pp. 1-14
DOI: 10.30636/jbpa.81.404

Making Null Results Credible: An Overview of Design and Analytical Tools

Nathan Favero*, Paolo Belardinelli†, Ling Zhu, Joanna Lahey††**

Abstract: Despite widespread acknowledgement of the importance of disseminating null results, researchers often struggle to successfully publish null findings. One common criticism leveled against such findings is that null results could be driven by design factors like inadequate sample size or measurement error. In this essay, we provide an overview of several tools and practices that researchers can implement, both during the design stage and during post-hoc analysis, to make null results more credible. Specifically, as researchers design their studies, they can make use of power analysis and preregistration, while taking care to follow best practices for variable measurement and—in the case of experimental studies—manipulation checks. During the analysis stage, researchers can move beyond "failing to reject the null hypothesis" by using confidence intervals, equivalence tests (such as the two one-sided tests (TOST) procedure), or Bayesian statistical approaches such as the Bayes Factor.

Keywords: Null findings, Research design, Power analysis, Preregistration, Equivalence

Social scientists have long recognized that null results are important and have even developed theories that predict null relationships (Rainey, 2014). If only statistically significant results appear in our journals, we risk creating a skewed perception of social reality. For example, suppose that a dozen studies have estimated the effect of sending SMS reminders on citizens' likelihood of paying a fine on time. If half of these studies find positive effects and half find null effects but the null results are never published, the published scholarly record may convey to policymakers and scientists alike an overstated optimism about the potential for SMS reminders to generate socially desirable outcomes.[1] In the context of meta-analytical methods, scholars often discuss how the existence of unpublished null results sitting in the proverbial "file drawer" can bias meta-analytical estimates of effect sizes. Formal methods for assessing publication bias have been developed and are widely used in meta-analyses. However, it is unclear whether we can reliably adjust estimates for publication bias, given limitations of such methods (Carter et al., 2019). Thus, scholars have argued that a better solution is to track down results from the "grey literature" of unpublished studies to create better meta-analytical estimates (Ringquist, 2013; Page et al., 2021). Despite this recognition of the need to make null results part of the scholarly record, null results are famously difficult to publish, and tools that aid in evaluating null results—such as equivalence testing and the Bayes factor—remain underutilized.

"Null results" generally refer to studies in which one fails to reject a null hypothesis of no effect for the key relationship(s) being examined.[2] By contrast, "positive results" refer to studies where a null hypothesis of no effect is rejected. While we join many other scholars in arguing that null results can be important, we also acknowledge that the informational value of a study with null results can vary substantially. Just as a study with positive findings will carry more weight if it is well-designed, a study with null findings will prove more informative if it is well-designed.

* American University, † Indiana University, ** University of Houston, †† Texas A&M University
Address correspondence to Nathan Favero at favero@american.edu.

Beyond general design issues, the ideal null result is one that is precisely estimated, with a small standard error (often shown as a narrow confidence interval around the point estimate). Unlike precise nulls, a null result based on an estimate with a very large standard error tells us little about the effect or association of interest.[3] Put differently, one reason results can be null is that the sample size is too small to let us conclude much of anything empirically.

More broadly, there are several reasons that one might find null results. The first is that the true effect size or association of interest may have a near-zero value. We refer to this as a "true null," meaning that the null hypothesis of no effect/association is true (or at least approximately true, such as when a true effect size is non-zero but negligible in size).

As already alluded to, however, a true null is not the only reason one might find a null result. Another common reason for null results is a lack of statistical power (due to too small a sample size, too little variation in the independent variable(s), and/or too much unexplained variation in the dependent variable). This is perhaps the most widely acknowledged reason that null results might not reflect a "true null." Yet there is an ongoing problem of authors treating null results as evidence of no effect without adequately considering whether estimates are precise enough to rule out effects of a substantial size (Rainey, 2014; Fitzgerald, 2025). Thus, null findings are often misinterpreted due to an incomplete appreciation for the fact that they can be driven by insufficient statistical power.

Potentially misleading null results may also occur due to a number of research design issues, such as poor measurement of key variables or failure of an experimental treatment to succeed in manipulating the independent variable of interest. Kane (2025) provides a nice inventory of reasons that null effects may occur within experimental designs. Observational research is equally vulnerable to null effects driven by poor research design, which can obscure substantively meaningful relationships among variables. Null results can also occur because of poor modeling choices (e.g., inefficient estimators) or bad luck (even well-powered studies occasionally fail to detect effects).

We also note that the importance of null results will depend on the importance of the research question being asked, just as with positive results. There are many unrelated variables whose lack of a relationship is substantively uninteresting. Thus, researchers disseminating null results should be deliberate in explaining why their research question and its null finding are important. While assessing the importance of research questions requires subjectivity, perceived importance nonetheless drives many decisions about how to prioritize attention and resources in the scientific community. We do not argue that by embracing the (potential) importance of null results, we should ignore other criteria for judging the importance of research questions. At the same time, we acknowledge the practical reality that one barrier to publishing null results is that reviewers and editors may be less likely to judge a research question as important if a study's hypotheses ultimately go unsupported empirically. Whether this reasoning has any legitimate basis is perhaps a matter of controversy (e.g., one might reasonably argue that an unsupported theory is less important than a supported theory). But certainly, many null results are important—particularly ones that are credibly estimated with reasonable precision and help answer a research question of scientific or practical importance.

The aim of this paper is to consider how researchers can best evaluate, interpret, and explain results that are potentially important but statistically insignificant. We review a number of practical tools that researchers can use both at the pre-hoc design stage and post-hoc in order to make their results more credible and their analyses more rigorous. While many of the practices we outline can be beneficial to any study—including ones with positive results—our particular focus is on how these tools may be useful in applications where results turn out to be null.

A Foundational Concept: Smallest Effect Size of Interest (SESOI)

Before discussing specific tools for null results, we first introduce an important concept we will repeatedly reference: the smallest effect size of interest (SESOI). Underlying this concept is the notion that what we care about is not just whether an effect size or association is non-zero but rather whether it is large enough to be considered substantively meaningful. Assessing whether an effect of a given magnitude should be considered "meaningful" as opposed to "negligible" (using the language of Rainey, 2014) requires subjective judgement, and credible evaluation will typically require drawing on subject-matter expertise (see Anderson, 2019). To facilitate precision and formal testing, it can be helpful to identify a specific threshold as the smallest effect size of interest (Lakens, 2017). For example, with a dependent variable that is a proportion, one might argue that any effect (positive or negative) with a magnitude of at least five percentage points is considered meaningful, while anything smaller should be considered insubstantial. Or, researchers can use standardized effect sizes, "small," "medium," and "large" from Cohen (1992) (e.g., Cohen's *d* for differences in means, Cohen's *w* for χ^2) or based on prior studies. For example, Cohen's *d* is the difference of means divided by the standard deviation, and he defines small, medium, and large effect sizes as $d < 0.20$, 0.50, and 0.80, respectively, based on psychology studies. Matthay et al. (2021) provide similar analytic formulas and small/medium/large effect sizes for additional tests (e.g., Relative Risk). Highhouse and Brooks (2023) argue, however, that benchmarks for effect sizes have generally been set too high (by Cohen), at least for studies of behavioral outcomes; more realistic benchmarks might be found from meta-analyses of existing studies, and Lenth (2001) argues that Cohen's *d* ignores precision of measurement. More broadly, the number of standard deviations apart may not be meaningful for policy makers who prefer an effect size that they consider to be small, medium, or large from a policy standpoint (e.g., dollars, lives saved). A key challenge in drawing conclusions of meaningful null effects is persuading the reader that one has reasonably identified a threshold for small or near-zero effects (for further advice, see Peetz et al., 2024).

Power Analysis

We first turn to design choices that can increase believability, should researchers find null results. Statistical power refers to the probability of correctly rejecting a false null hypothesis—that is, the probability of detecting an effect when the effect exists.[4] Underpowered null results may be insignificant, not because there is no true effect, but because the sample size is too small to detect the true effect. Of course, concerns related to null findings are not the only reason that power is important. Studies with low power also are more likely to provide inflated estimated effect sizes and to have wrong-signed estimates (Bagilet, 2024; Black et al., 2022; Campos-Mercade, 2024; Doucette, 2025; Egerod & Hollenbach, 2024; Gelman & Carlin, 2014; Kaestner, 2021; Matthay & Glymour, 2022). But for the purposes of this paper, we are interested in power's relationship to null findings. Analytic and simulated methods that measure power can help interpret null results.

Power analysis can be used a priori in experiments (or other data collection) to determine the minimum sample size needed for the experiment to be "fully powered," thus lending more credibility to null findings. Power analysis is also useful in non-experimental work. Power calculations or simulations can determine what size of effects are powered in previously collected datasets for a given level of significance (alpha). For example, a dataset and analysis method may be only powered to find large effects, meaning that a null result could be hiding a true small or medium effect, but is less likely to be hiding a true large effect. A different dataset or analysis method could be powered to find small effects, making a null result more meaningful. Additionally, a null finding when the measured power is 80% or higher (for the effect size of interest) is more meaningful than a null finding with less power (Campos-Mercade, 2024). While power analysis can be done on data that have already been collected, it is important to use a priori theorized effect sizes, not the estimated effect sizes from data analysis (Black et al., 2022; Hoenig & Heisey, 2001; Lenth, 2001).

Power analysis can be done through closed-form analytic methods or via simulation. Analytic power analysis is often used for experiments with simple empirical designs. To use an analytic method, assumptions need to be made about power, statistical significance, and effect sizes. Generally, researchers assume a power of 80% or more and an alpha of 0.05, although some journals, such as *Nature Human Behavior,* now require a power of 95% or more.[5]

After determining the appropriate effect size, using approaches discussed in the earlier SESOI section, SESOI measured via policy-relevant units can be converted into Cohen's *d* values in order to use an analytic sample size program. Having a large enough sample to be powered to detect the smallest policy-relevant effect size increases the relevance and credibility of null findings. Several programs help with finding analytic solutions to sample size calculations. G*Power (Faul et al., 2007) is a popular free program focused on power analysis. Other popular programs include the *power* command in Stata and *pwr* in R (Champeley et al., 2020).

Interaction effects deserve additional attention. Simply doubling the necessary sample size when trying to determine an interaction effect results in underpowered results. Scholars generally recommend heuristics of four to sixteen times the sample size to obtain a fully powered interaction effect (Gelman, 2018; Baranger et al., 2023). Baranger et al. (2023) provide the R package *InteractionPoweR* and detailed information on how to incorporate interaction effects in sample size calculations.

As empirical analyses become more complicated, analytic methods become less tractable, and, in some cases, do not have closed-form solutions. Simulated power analysis can be used to measure the power for difference-in-differences (DiD) (Bagilet, 2024), staggered DiD (Egerod & Hollenbach, 2024), instrumental variables (Bagilet, 2024), and other commonly used analyses (Bagilet, 2024).

There are several guides to doing power analyses via simulations (Bagilet, 2024; Campos-Mercade, 2024 is especially user-friendly; Luedicke, 2013). Several user-generated programs also aid researchers with simulations. Bagilet (2024) provides simulation code for R*,* Luedicke (2013) provides *powersim* and Burlig et al. (2020) provides *pcpanel* for Stata.[6]

The general idea is as follows:

1. Either generate synthetic data based on a distribution and specific parameters or use real data.
2. Randomly manipulate the data to match a "true" hypothesized treatment effect (ex., randomly assign treatment = 1 based on the model and hypothesized alpha, randomly assign treatment outcome to those in the treatment group).
3. Run your model on the data, record p-value (or estimate of interest and standard error).
4. Repeat 1-3 for a pre-specified number of repetitions (ex., 1,000).
5. Power will be the % of p-values recorded in #3 that are less than your specified alpha (ex., if alpha is 0.05, and 800 of the 1000 repetitions have p<0.05, then power will be 80%).

Pre-registration

Another pre-hoc procedure that researchers can do to make their findings, including null findings, more believable is to pre-register their analyses. Pre-registration involves storing a document that details a research plan in a public repository before the empirical data are collected by the researcher(s). The document usually specifies the hypotheses to be tested and includes information about how the data will be collected and analyzed. Pre-registration encourages researchers to focus more on sound theory and methodology rather than on the statistical significance of results, thereby improving the quality of studies regardless of the statistical outcomes. As we detail below, it can increase the value of studies that find null results.

The practice of pre-registering empirical studies, in particular experiments, has gained popularity across disciplines over time, starting from psychology (Strømland, 2019; van den Akker et al., 2024), followed by economics (Banerjee et al., 2020; Olken, 2015), management (Toth et al., 2021), political science (Monogan, 2015), and sociology (Manago, 2023). Public administration recently joined this movement (Belardinelli & Zhu, 2025).

Pre-registration allows researchers to distinguish confirmatory analysis from exploratory analysis (Nosek et al., 2018). Predicted findings are considered more robust than exploratory findings because they stem from prior theory and hypotheses established before data collection and are therefore more likely to reflect underlying general rules that can be replicated (Sarafoglou et al., 2022; van't Veer & Giner-Sorolla, 2016; Wagenmakers et al., 2012). In contrast, exploratory findings that arise from post hoc analyses without predefined hypotheses are more susceptible to multiple hypothesis testing concerns, false positives, and p-hacking with *ex post* hypotheses developed to fit findings.

Pre-registration fosters a credible commitment by researchers to a pre-established set of analyses and therefore reduces the probability of reporting false positives resulting from multiple tests, which is particularly relevant when reported results are significant. However, pre-registration enhances the credibility of both null and non-null findings by encouraging researchers to think more carefully about theoretical predictions and methodological choices before data are collected and analyzed. As a result, in pre-registered studies, null findings are less likely to stem from poor design or power issues.

Clearly, researchers may still hesitate to submit, and reviewers to support or editors to publish, studies that fail to confirm their pre-registered hypotheses. This is the so-called 'file drawer problem' (Rosenthal, 1979). In this respect, pre-registration alone cannot fully solve the problem of publication bias. However, there is evidence that pre-registered studies are more likely to report null findings than non-pre-registered ones (Belardinelli & Zhu, 2025; Toth et al., 2021; though c.f. van den Akker et al., 2024), possibly because pre-registered studies are generally seen as more credible, even when reporting null findings. Ultimately, pre-registration shifts the focus away from the results of individual studies toward thinking in terms of expected values, replication, and the mitigation of publication bias. In this way, null findings become just as important as positive findings.

Whereas pre-registration is mostly adopted for experiments (Banerjee et al., 2020; Ofosu and Posner, 2021), the benefits of this practice can also be realized for non-experimental studies. Clearly, the latter involves more challenges, including the need to demonstrate that the hypotheses and analysis plan were developed before accessing the data. This is feasible in observational contexts where researchers collect their own data (either quantitative or qualitative; Jacobs, 2020), in prospective studies where data will be observed in the future (e.g., Neumark, 2001), or when using restricted-access data (Burlig, 2018).

Belardinelli and Zhu (2025) discuss three conditions for pre-registrations to generate their intended benefits: (i) they should be stored before the data are observed; (ii) published pre-registered studies should reflect the execution of the preregistered plans as closely as possible; and (iii) deviations from the original plans should be transparently reported in the published studies. Nevertheless, these conditions take for granted that each pre-registration is properly done, largely overlooking the elements that enhance the quality of an individual pre-registration.

Several researchers across the social sciences have developed guidelines and checklists to help develop a good pre-registration (e.g., Banerjee et al., 2020; Chen & Grady, 2019; McKenzie, 2012; Van't Veer & Giner-Sorolla, 2016), and reporting standards have been established for different types of studies, such as SPIRIT for trials (Moher and Chan, 2014) and PRISMA-P for systematic reviews and meta-analyses (Shamseer et al., 2015). While standards for pre-registration differ across fields and are still being debated, we suggest these essential elements:

1. *Hypotheses* – These should specify the outcomes of interest (dependent variables), the explanatory independent variables, any possible moderating or mediating variables, and the direction of the hypothesized effects.
2. *Data collection procedure* – For secondary data, it is important to describe the sources and how they were obtained. When collecting primary data, the data collection process should be described, along with the research instruments used, such as surveys and/or the experimental design.
3. *Measures* – It is important to report how each variable will be operationalized and measured. In the case of randomized experiments, it is also important to detail the treatments and their levels.
4. *Sample* – The pre-registration should report information on the expected sample size and its rationale, including power analysis (see the previous section of this paper) and SESOI.
5. *Statistical analysis* – The statistical models adopted to test the hypotheses should be specified in the pre-registration (for example, OLS, ANOVA, logistic regression models).

Once data are collected, it is important for researchers to specify that the study was preregistered, to clearly distinguish preregistered and exploratory analysis, and to transparently report – and justify – any deviations from the original plan. Such deviations may arise for several reasons, which Lakens (2024) groups into five categories, each with its own implications for the validity of the results: unforeseen events, mistakes in the preregistration, missing information, violations of assumptions, and falsification of auxiliary hypotheses. Clearly, null findings are not a valid reason for deviating from preregistration. If preregistered hypotheses and analyses yield null results, these should still be included in the study.

The elements outlined here may oversimplify and overlook other relevant aspects. For example, pre-registering a large number of hypotheses can undermine the value of pre-registration, especially when there is no distinction between core and secondary hypotheses (Banerjee et al., 2020) and no statistical correction for multiple tests during data analysis (EGAP website, 2025; see also García-Pérez, 2023). Additionally, in the context of statistical analysis, decisions about how to handle missing data, outliers, attrition, and similar issues can significantly affect the results. While awareness of such challenges is certainly important, the required fields of the main platforms used for pre-registering empirical studies (such as the Open Science Framework Registry or the American Economic Association RCT Registry[7]) cover the elements outlined above and provide a foundation for a high-quality pre-registration. The peer-review process would still be in place to address any remaining gaps.

Measurement Error and Manipulation Checks

As noted previously, one potential source of null results is measurement error. While measurement error can also sometimes cause false positive results, measurement error can easily lead to null results since noisy data generally makes it more difficult to detect underlying associations. Stated more formally, random measurement error often leads to attenuation bias (biasing of associations toward zero) (McAdams, 1986; Jackman, 2008).

Several design factors can help mitigate potential concerns about measurement error (Kane, 2025). When measuring attitudes or complex behaviors through surveys, using multi-item measures is considered best practice, since these measures often allow for more precise measurement than single-item variables. In some cases, especially when well-established measures of key variables are unavailable, it may also be helpful to measure secondary variables that are expected to correlate strongly with key variables of interest. This will allow for tests of criterion validity, in which one verifies whether key variables of interest yield the expected correlations with related variables. For example, showing readers that the dependent variable has an as expected significant relationship with one or more secondary variables (alongside the null result for the main independent variable of interest) can build confidence in the validity of the dependent variable's measurement (and perhaps of the data collection more broadly).

Another tool for considering measurement error in surveys is attention checks, which refer to survey questions that test for general attentiveness on a survey. A common example is a survey item where the prompt explicitly instructs respondents to select a particular answer (e.g., "Please select strongly agree."). Inattentive respondents are likely to provide unreliable survey responses, subject to measurement error.

Manipulation checks for experimental studies

For experimental work, a key concern is whether experimental manipulation(s) will succeed in altering the key independent variable(s) in the intended manner. Manipulation checks[8] involve measuring whether respondents have absorbed the experimental manipulation in the expected manner. For example, factual manipulation checks involve asking respondents to recall factual information that differed across treatment conditions in order to see whether they were attentive to the key information that was part of a treatment (Kane & Barabas, 2019). In cases where a clear distinction can be drawn between the manipulation itself and the underlying variable one is attempting to manipulate (e.g., anger), it is good to directly measure the underlying variable and see whether its value differs by treatment condition (Mutz, 2021). Researchers debate whether to administer manipulation checks prior to measurement of the dependent variable (Kane & Barabas, 2019), but including such checks after the dependent variable is least likely to raise concerns (Mutz, 2021).

Data created through manipulation checks provides informational value, since it can help to narrow down potential explanations for why a particular result for the dependent variable was found. However, it is generally not appropriate to drop observations from respondents who fail manipulation checks since internal validity can be compromised if attrition patterns differ across experimental groups (Mutz, 2021). In contrast, screening out respondents based on pre-treatment general attention checks can be appropriate, particularly when the dropping protocol has been pre-registered.

In general, reducing measurement error and providing experimental manipulation checks should be considered as necessary, but not sufficient, for the credibility of null results. Researchers should still make sure their design is sufficiently powered and consider pre-registration and/or the techniques discussed below.

Confidence Intervals

Our first post-hoc methodology to lend credibility to null effects is the simple confidence interval. Suppose a researcher estimates a relationship between variables and finds that all values within the resulting 95% confidence interval are smaller in magnitude than the SESOI. This would suggest that the researcher has grounds to conclude that there is, at most, only a slight (near-zero, or policy-irrelevant) relationship.

Confidence intervals will be narrower (making it easier to find precise null results) when an efficient estimator is utilized. Thus, one important analytical step is trying to ensure one uses the best estimator, considering not only bias but also efficiency. For example, controlling for strong predictors of the dependent variable often benefits efficiency of estimation, even if there is no concern about confounding effects (Mutz et al., 2019); such modeling choices would ideally be pre-registered.

For models where confidence intervals are not readily available through analytical computation for substantive effect size estimates, such intervals can be created through simulation (Rainey 2014).

TOST

A more formal approach to using confidence intervals in the manner described above is known as the two one-sided tests (TOST) procedure—a type of equivalence test (Rainey, 2014; Lakens, 2017). One must first identify the SESOI. Then, for a test at the 0.05 alpha level, one would construct a 90% confidence interval (the 90% level is somewhat counterintuitive, but this is because a two-sided confidence interval is being used to conduct one-sided tests). If the confidence interval doesn't contain any values with an absolute value greater than the SESOI, one concludes that there is evidence of equivalence (a negligible effect).

While use of confidence intervals for TOST is sound, one can also obtain precise p-values for an equivalence test, using readily available packages in mainstream statistical software (e.g., *TOSTER* in R (Lakens & Caldwell, 2025), *tost* in Stata (Dinno, 2024)). It is also possible at the design stage to conduct power analysis for TOST, ensuring the sample size is large enough to conclude a negligible effect under the assumption of a true null (Lakens, 2017).

We note that there is some potential overlap between power analysis and the use of equivalence testing. Indeed, early practice for establishing evidence of equivalence or near-zero effects sometimes took the form of what has been called the "power approach," which consisted of checking for non-significance alongside high statistical power (see Meyners, 2012). More recent guidance, however, typically recommends making use of formal equivalence tests such as TOST or Bayesian approaches (Hoenig & Heisey, 2001; Lakens, 2017; Fitzgerald, 2025).

Bayesian Hypothesis Testing and the Bayes Factor

Bayesian hypothesis testing provides another alternative approach for evaluating hypotheses. It uses the probability of parameters of interest to evaluate the evidence for or against a hypothesis. The classic null hypothesis significance test (NHST) relies on the use of a p-value, defined as the probability of a test statistic that is at least as extreme as the observed one under the null hypothesis. In contrast to NHST, Bayesian hypothesis testing directly evaluates the probability of each hypothesis given the observed data. The Bayesian approach also allows flexibility in formulating the null hypothesis, such as defining a null region rather than a single null quantity (i.e., an effect of zero).

A key feature of the Bayesian approach is the incorporation of prior knowledge through prior distributions for the quantity of interest. In practice, one can specify a prior distribution of a given parameter value using published results from the existing literature (see, for example, Gill & Witko, 2013). Referred to as informative priors, these prior specifications may influence the estimation of the posterior distribution, especially when data are limited. One can also choose to use an uninformative prior (such as the uniform distribution), which will have minimal influence on the posterior distribution.[9] The Bayesian approach produces full posterior distributions of parameter values based on observed data and prior distributions by the following steps (for examples of how to use and report Bayesian hypothesis testing, see Fienberg, 2011; Gill and Witko, 2013; Dienes, 2021).

1. For a specific hypothesis, define the prior distribution of the parameter of interest.
2. Derive the posterior distribution of parameter values from data.
3. Construct credible intervals.
4. Accept (or reject) the point-specific hypothesis (e.g., a hypothesis of null effects) based on the distributional quantiles and the probability statements about the posterior distributions (Gill & Witko, 2013).

For example, if one were to test the effect of SMS nudging on citizens' likelihood of paying a fine on time using Bayesian hypothesis testing, one could summarize a posterior distribution and interpret it as the probability that the parameter of interest is less than or greater than a specific value. In other words, if a null interval can be determined, when a proportion of the posterior distribution (e.g., a 95% credible interval) falls in the null interval, then it supports the null hypothesis that the SMS nudging does not have any effect on changing citizen behavior (see Harms & Lakens, 2018).

Bayesian hypothesis testing can also be applied to compare competing and alternative hypotheses, which involves the use of the Bayes Factor (Kass & Raftery, 1995). Proposed by Jefferys (1935, 1961), the Bayes Factor (BF) is defined as the ratio of the integrated marginal likelihoods of two competing models. Using BF, one can calculate the posterior odds under the null hypothesis and the posterior odds under the alternative hypothesis. As such, BF can be interpreted as measuring the relative evidence for the two competing hypotheses (H_0 and H_a), in other words, the relative success of the two competing hypotheses at predicting the observed data.[10]

With the abovementioned example of SMS nudging, suppose a researcher is to test two competing hypotheses about the nudging effect, described as a parameter value b, and consider the SESOI to be 5 percentage points. Suppose the null intervals are determined as $[0,5]$. As such, the null hypothesis is that $b_{HO} \in [0,5]$, and the competing hypothesis is that $\mathbf{b_{Ha}<0}$ or $\mathbf{b_{Ha}>5}$. In other words, the null hypothesis expects a small positive nudging effect (greater than 0 but less than 5 percentage points), and the alternative hypothesis posits that the nudging effect is either greater than 5 or smaller than 0. To compute BF, one will start by defining two prior distributions for b_{H0} and $b_{Ha,}$ expressing the probability of the value of b under the two hypotheses, given data, **D:** $\mathbf{p_{H0} = p(b \in [0,5] | D)}$ and $\mathbf{p_{Ha} = p(b <0 \text{ or } b>5 | D)}$. Next, using the same data and likelihood function, two posterior distributions are derived for the above-mentioned two hypotheses: $\pi_{H0} = p(b = b_{H0})$ and $\pi_{Ha} = p(b = b_{Ha})$. As Gill and Witko (2013, p.10) summarize, BF is given as the "posterior odds over the prior odds": $(\frac{\pi_{H0}}{\pi_{H0}})/(\frac{P_{H0}}{P_{Ha}})$.

The estimation of posterior distributions is implemented by using Markov Chain Monte Carlo algorithms. Various statistical software provide estimation commands for Bayesian analysis and Bayesian hypothesis testing. For example, *Stan,* a widely used platform for Bayesian analysis, has interfaces for both R and Python, through *RStan* and *PyStan*, respectively. R has several additional packages supporting Bayesian analysis (e.g., package *rjags*) and Bayesian hypothesis testing (e.g., package *BayesFactor*). Stata has a suite of Bayesian commands, including the *bayestest* command for Bayesian hypothesis testing and the *bayesstats ic* command for computing BF.

BF is a more intuitive way of evaluating evidence for the null hypothesis compared to evidence for the alternative hypothesis. It addresses many concerns around NHST and the use of p-values in hypothesis testing. First, BF is directly constructed based on observed data rather than assumptions about the sample and population. Second, using BF, researchers can draw an explicit conclusion about the plausibility of the null hypothesis being true. In contrast, when using NHST, we never directly find evidence that the null hypothesis is true, only failure to reject it. Moreover, BF directs our attention to the strength of evidence for a null effect rather than artificially chosen levels of statistical significance (Morey & Rounder, 2011; Neely, 2019). As BF is calculated as a ratio, a value greater than 1 indicates there is more evidence supporting the null hypothesis over the alternative hypothesis. In addition, BF does not directly incorporate the sample size in its calculation.

Despite its advantages, there are several caveats and limitations associated with Bayesian hypothesis testing and the use of BF. Because the choice of prior distributions is crucial to the estimation process, the computation of BF can be sensitive to the subjective choice of prior specifications. There is also a potential for computational difficulty when one constructs complex models. Because it involves integrating marginal likelihoods, computing BF can be time-consuming or lead to model nonconvergence issues when one specifies a model with a large number of parameters and/or hierarchical parameter structures.

Conclusion

Credible null results are important for the progression of knowledge. This paper discusses different methods that researchers can use both during the design process and after data have been collected to make null results more believable, and thus, hopefully, more publishable. Power analysis and pre-registration can increase confidence that null results are not the result of low sample size or running atheoretical analyses. Careful experimental design, including attention and manipulation checks, reduces measurement error and helps determine that the experimental manipulation was not ignored. A small toolbox of post-hoc methods also serves to enhance the credibility of null results. A standard 95% confidence interval can sometimes rule out policy-relevant effects. Similarly, an equivalence test called the two one-sided tests (TOST) procedure is another way to rule out large effect sizes. Finally, the Bayes Factor can provide evidence directly supporting a null hypothesis, rather than simply failing to reject it as with standard null hypothesis testing.

While the methods we suggest are primarily for researchers to use, we hope that our discussion has also helped readers better understand what makes null results more trustworthy, and thus, publishable. There are additional systemic steps that could be taken to make it easier or more rewarding for researchers in pursuit of knowledge. For example, committing to publish a final paper based on its pre-registration plan (as with Neumark, 2001) or coordinating multiple labs to work on problems in which detecting a small but practically relevant effect size requires sample sizes that exceed what one team of researchers could practically recruit (Klein et al., 2014). We hope that researchers, reviewers, and editors alike find this overview of tools, papers, and programs we cite useful for identifying convincing evidence of null findings.

Notes

1. To be precise, the literature in this hypothetical example is likely to offer a misleading picture if the unpublished studies tend to have different (smaller) effect size estimates than the published studies; small effect size estimates often lead to null results, but results can also be null because of inadequate power.
2. While the concept of null results emerges out of null hypothesis significance testing (a framework that has been widely criticized; see Gill, 1999; Gill & Meier, 2000), a similar issue can occur under alternative frameworks. More broadly (and informally), “null results” might be understood to describe studies where results indicate that the effect/association of interest may well be zero (e.g., credible intervals that contain zero under a Bayesian framework).
3. There still may be information value from an underpowered study in terms of what the study can tell us about the potential for a particular research design or dataset to yield precise estimates, since things like random attrition rates or natural variance that affect statistical power may have been difficult to anticipate.
4. This latter explanation assumes the null hypothesis indicates no effect—typical practice in most of social science.
5. See https://www.nature.com/nathumbehav/submission-guidelines/registeredreports . *Last accessed December 3, 2025.*
6. Stata also provides generalized instructions for using their *simulate* and *power* commands together for simulated power analysis. https://www.stata.com/support/faqs/statistics/power-by-simulation/ *Last accessed 6/4/25.*
7. See https://osf.io/ and https://www.socialscienceregistry.org/ respectively. *Last accessed 8/18/25.*

8. Note that manipulation checks are not the same as balance tests, which check to see how similar the treatment and control groups are. Balance tests are usually more of a concern for spurious significant results when differences are actually attributable to observable differences between the treatment and control groups rather than the treatment itself. It is possible that group imbalances could cancel out real effects, leading to spurious null results. For more on the perils of balance testing, see Mutz et al. (2019).
9. Bland (2025) compares how informative and uninformative priors influence the estimation of posterior distributions, with examples of Bayesian structural models in economic experiments.
10. There has been a steady increase in the use of Bayes Factor as a tool for hypothesis testing in social science research; for examples, see Hoijtink et al. (2019).

References

Anderson, A. A. (2019). Assessing statistical re sults: Magnitude, precision, and model uncertainty. *The American Statistician,* 73(sup1), 118-121. https://doi.org/10.1080/00031305.2018.1537889

Bagilet, V. (2024). Accurate Estimation of Small Effects: Illustration Through Air Pollution and Health. *Working Paper.* https://vincentbagilet.github.io/inference_pollution/

Banerjee, A., Duflo, E., Finkelstein, A., Katz, L. F., Olken, B. A., & Sautmann, A. (2020). In praise of moderation: Suggestions for the scope and use of pre-analysis plans for RCTs in economics (No. w26993). *National Bureau of Economic Research.* https://doi.org/10.3386/w26993

Baranger, D. A. A., Finsaas, M. C., Goldstein, B. L., Vize, C. E., Lynam, D. R., & Olino, T. M. (2023). Tutorial: Power Analyses for Interaction Effects in Cross-Sectional Regressions. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231187531. https://doi.org/10.1177/25152459231187531

Belardinelli, P., & Zhu, X. (2025). Pre-Registering Public Administration Studies: Avoiding the Poor Practice of a 'Best-Practice'. *Journal of Behavioral Public Administration*, *8*, 1-14. https://doi.org/10.30636/jbpa.81.377

Black, B., Hollingsworth, A., Nunes, L., & Simon, K. (2022). Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion. *Journal of Public Economics*, *213*(C),104713. https://doi.org/10.1016/j.jpubeco.2022.104713

Bland, J. (2025). Some guidance for the choice of priors for Bayesian structural models in economic experiments. *Journal of the Economic Science Association*, 1-10. https://doi.org/10.1017/esa.2025.6

Burlig, F. (2018). Improving transparency in ob servational social science research: A pre-analysis plan approach. *Economics Letters*, *168*, 56-60. https://doi.org/10.1016/j.econlet.2018.03.036

Burlig, F., Preonas, L., & Woerman, M. (2020). Panel data and experimental design. *Journal of Development Economics*, *144*, 102458. https://doi.org/10.1016/j.jdeveco.2020.102458

Campos-Mercade, P. (2024). *Power analysis through simulations in Stata: A step-by-step guide.Working Paper.* https://www.econstor.eu/handle/10419/302273

Carter EC, Schönbrodt FD, Gervais WM, Hilgard J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*. *2*(2),115-144. https://doi.org/10.1177/2515245919847196

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H. (2020). *pwr: Basic Functions for Power Analysis.* In [Computer software]. Comprehensive R Archive Network (CRAN).

https://CRAN.R-project.org/package=pwr

Chen, L., & Grady, C. (2019). 10 Things to Know about Pre-analysis Plans. *Evidence in Governance and Politics (EGAP)*, Institute of Government Studies, University of California, Berkeley. Link: https://egap.org/resource/10-things-to-know-about-pre-analysis-plans/

Cohen, J. (1992). A power primer. *Psychol Bull*, *112*(1), 155-159. https://doi.org/10.1037//0033-2909.112.1.155

Dienes, Z. (2021). How to Use and Report Bayes ian Hypothesis Tests. *Psychology of Consciousness: Theory, Research, and Practice*, *8*(1), 9–26. https://doi.org/10.1037/cns0000258

Dinno, A. (2024). TOST: Two One-Sided Tests of Equivalence (Version 3.1.5) [Stata package]. Portland State University. Available from https://www.alexisdinno.com/stata/tost.html

Doucette, J. S. (2025). What Can We Learn about the Effects of Democracy Using Cross-National Data? *American Political Science Review*, 199(3), 1549-1558 https://doi.org/10.1017/S0003055424001278

EGAP website. (2025). *10 Things You Need to Know About Multiple Comparisons*. Link: https://methods.egap.org/guides/analysis-procedures/multiple-comparisons_en.html (Last accessed: November 24, 2025).

Egerod, B. C., & Hollenbach, F. M. (2024). How many is enough? Sample Size in Staggered Difference-in-Differences Designs. *OSF Preprint*. https://doi.org/10.31219/osf.io/ac5ru

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. *Behavioral Research Methods*, *39*(2), 175-191. https://doi.org/10.3758/bf03193146

Fienberg, S. E. (2011). Bayesian Models and Meth ods in Public Policy and Government Settings. *Statistical Science*, 26(2), 212-226. https://doi.org/10.1214/10-STS331

Fitzgerald, J. 2025. The Need for Equivalence Testing in Economics. MetaArXiv. https://osf.io/preprints/metaarxiv/d7sqr_v1

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. (2018). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, *44*(1), 16-23. https://doi.org/10.1177/0146167217729162

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641-651. https://doi.org/10.1177/1745691614551642

Gill, J. (1999). The Insignificance of Null Hypothe sis Significance Testing. *Political Research Quarterly,* 52(3), 647-674. https://doi.org/10.1177/106591299905200309

Gill, J., & K. J. Meier. (2000). Public Administra tion Research and Practice: A Methodological Manifesto. *Journal of Public Administration Research and Theory, 10*(1), 157-199. https://doi.org/10.1093/oxfordjournals.jpart.a024262

Gill, J. & C. Witko. (2013). Bayesian Analytical Methods: A Methodological Prescription for Public Administration. *Journal of Public Administration Research and Theory, 23*(2), 457-494. https://doi.org/10.1093/jopart/mus091

Harms, C., & Lakens, D. (2018). Making 'null ef fects' informative: statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research, 3*(S2): 382-393. http://dx.doi.org/10.18053/jctres.03.2017S2.007

Highhouse, S., & Brooks, M. E. (2023). Interpret ing the magnitude of predictor effect sizes: It is time for more sensible benchmarks. *Industrial and Organizational Psychology, 16*(3), 332-335. https://doi.org/10.1017/iop.2023.30

Hoenig, J. M., & Heisey, D. M. (2001). The Abuse of Power. *The American Statistician*, *55*(1), 19-24. https://doi.org/10.1198/000313001300339897

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539-556. https://doi.org/10.1037/met0000201

Jackman, S. (2008). Measurement. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford Handbook of Political Methodology* (pp. 119-151). New York: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199286546.003.0006

Jacobs, A. M. (2020). Pre-registration and results-free review in observational and qualitative research. In C. Elman, J. Mahoney, & J. Gerring (Eds.), *The production of knowledge: Enhancing progress in social science* (pp. 221–264). Cambridge University Press. https://doi.org/10.1017/9781108762519.009

Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*(2), 203-222. https://doi.org/10.1017/S030500410001330X

Jeffreys, H. (1961). *Theory of probability (3rd edt.).* Oxford University Press (Vol. 432).

Kaestner, R. (2021). Mortality and science: a com ment on two articles on the effects of health insurance on mortality. *Econ Journal Watch*, *18*(2), 192.

Kane, J. V., & J. Barabas. (2019). No harm in checking: Using factual manipulation checks to assess attentiveness in experiments. *American Journal of Political Science, 63*(1), 234-249. https://doi.org/10.1111/ajps.12396

Kane, J. V. (2025). More than meets the ITT: A guide for anticipating and investigating nonsignificant results in survey experiments. *Journal of Experimental Political Science, 12*(1), 110-125. https://doi.org/10.1017/XPS.2024.1

Kass, R. E. & A. E. Raftery. (1995). Bayes Factors. *Journal of the American Statistical Association,* 90(430), 773-785. https://doi.org/10.1080/01621459.1995.10476572

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355-362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2024). When and how to deviate from a preregistration. *Collabra: Psychology*, *10*(1), 117094. https://doi.org/10.1525/collabra.117094

Lakens, D., & Caldwell, A. (2025). TOSTER: Two One-Sided Tests (TOST) Equivalence Testing (Version 0.8.4) [R package]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=TOSTER

Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician*, *55*(3), 187-193. https://doi.org/10.1198/000313001317098149

Luedicke, J. (2013). Simulation-based power analy sis for linear and generalized linear models. *Stata Conference*, 1-25. https://ideas.repec.org/p/boc/norl13/13.html

Manago, B. (2023). Preregistration and registered reports in sociology: Strengths, weaknesses, and other considerations. *The American Sociologist*, *54*(1), 193-210. https://doi.org/10.1007/s12108-023-09563-6

Matthay, E. C., & Glymour, M. M. (2022). Causal Inference Challenges and New Directions for Epidemiologic Research on the Health Effects of Social Policies. *Current*

Epidemiology Reports, *9*(1), 22-37. https://doi.org/10.1007/s40471-022-00288-7

Matthay, E. C., Hagan, E., Gottlieb, L. M., Tan, M. L., Vlahov, D., Adler, N., & Glymour, M. M. (2021). Powering population health research: Considerations for plausible and actionable effect sizes. *SSM - Population Health*, *14*, 100789. https://doi.org/https://doi.org/10.1016/j.ssmph.2021.100789

McAdams, J. (1986). Alternatives for Dealing with Errors in the Variables: An Example Using Panel Data. *American Journal of Political Science,* 30(1), 256-278. https://doi.org/10.2307/2111304

McKenzie, D. (2012). A Pre-analysis Plan Checklist. *Development Impact, World Bank Blogs*. Link: https://blogs.worldbank.org/en/impactevaluations/a-pre-analysis-plan-checklist

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

Moher, D., & Chan, A. W. (2014). SPIRIT (standard protocol items: recommendations for interventional trials). *Guidelines for Reporting Health Research: a user's manual*, 56-67. https://doi.org/10.1002/9781118715598.ch7

Monogan III, J. E. (2015). Research preregistration in political science: The case, counterarguments, and a response to critiques. *PS: Political Science & Politics*, *48*(3), 425-429. https://doi.org/10.1017/S1049096515000189

Morey, R. D. & J. N. Rouder. (2011). Bayes Factor Approach for Testing Interval Null Hypotheses. *Psychological Methods*, *16*(4), 406-419. https://doi.org/10.1037/a0024377

Mutz, D. C., Pemantle, R., & Pham, P. (2019). The Perils of Balance Testing in Experimental Design: Messy Analyses of Clean Data. *The American Statistician*, 73(1), Article 1. https://doi.org/10.1080/00031305.2017.1322143

Mutz, D. C. (2021). Improving Experimental Treatments in Political Science. In Chapter. In J. N. Druckman & D. P. Green (Eds.), *Advances in Experimental Political Science* (pp. 219–38). Cambridge: Cambridge University Press.

Neely, S. R. (2019). Science V. Significance: Examining the Role and Application of Statistical Significance Testing in Public Administration Research. *Public Administration Quarterly*, *43*(2), 185-221. https://doi.org/10.1177/073491491904300202

Neumark, David. (2001). The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design. *Industrial Relations: A Journal of Economy and Society*, *40*(1), 121-44. https://doi.org/10.1111/0019-8676.00199

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), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Ofosu, G. K., & Posner, D. N. (2023). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, *21*(1), 174-190. https://doi.org/10.1017/S1537592721000931

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61-80. https://doi.org/10.1257/jep.29.3.61

Page, M. J., Sterne, J. A., Higgins, J. P., & Egger, M. (2021). Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods*, *12*(2), 248-259. https://doi.org/10.1002/jrsm.1468

Peetz, H. K., Primbs, M., Dudda, L., Andresen, P. K., Pennington, C. R., Westwood, S., & Buchanan, E. M. (2024). Evaluating interventions: A Practical Primer for Specifying the Smallest Effect Size of Interest. https://doi.org/10.31234/osf.io/3qmj4

Rainey, C. (2014). Arguing for a negligible effect. *American Journal of Political Science, 58*(4), 1083-1091. https://doi.org/10.1111/ajps.12102

Ringquist, E. (2013). *Meta-analysis for Public Management and Policy*. John Wiley & Sons.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638. https://doi.org/10.1037/0033-2909.86.3.638

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E. J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), 211997. https://doi.org/10.1098/rsos.211997

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj*, *349*, g7647. https://doi.org/10.1136/bmj.g7647

Strømland, E. (2019). Preregistration and reproducibility. *Journal of Economic Psychology*, 75, 102143. https://doi.org/10.1016/j.joep.2019.01.006

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J. & Borns, J. (2021). Study preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology*, *36*(4), 553-571. https://doi.org/10.1007/s10869-020-09695-3

van den Akker, O. R., van Assen, M. A., Bakker, M., Elsherif, M., Wong, T. K., & Wicherts, J. M. (2024). Preregistration in practice: A comparison of preregistered and non-preregistered studies in psychology. *Behavior Research Methods*, *56*(6), 5424-5433. https://doi.org/10.3758/s13428-023-02277-0

Van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12. https://doi.org/10.1016/j.jesp.2016.03.004

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632-638. https://doi.org/10.1177/1745691612463078