Beyond p-values: Utilizing Multiple Estimates to Evaluate Evidence

|  |
| --- |
| Kathrene D. Valentine1, Erin M. Buchanan2, John E. Scofield1, & Marshall Beauchamp3 |
|  |
| 1 University of Missouri |
| 2 Missouri State University |
| 3 University of Missouri - Kansas City |
|  |

Author note

Kathrene D. Valentine and John E. Scofield are Ph.D. candidates at the University of Missouri. Marshall Beauchamp is a Ph.D. candidate at the University of Missouri - Kansas City. Erin M. Buchanan is an Associate Professor of Quantitative Psychology at Missouri State University.

Correspondence concerning this article should be addressed to Kathrene D. Valentine, Postal address. E-mail: [my@email.com](mailto:my@email.com)

Abstract

Enter abstract here (note the indentation, if you start a new paragraph).

*Keywords:* keywords

Word count: X

Beyond p-values: Utilizing Multiple Estimates to Evaluate Evidence

Recent events in psychological science have prompted concerns within the discipline regarding research practices and ultimately the validity and reproducibility of published reports. One often discussed matter is over-reliance, abuse, and potential hacking of p-values produced by frequentist null hypothesis significance testing (NHST), as well misinterpretations of NHST results (Gigerenzer, 2004; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). We agree with these concerns and believe that many before us have voiced sound, generally accepted opinions on potential remedies, such as an increased focus on effect sizes (Cumming, 2008; Lakens, 2013; Maxwell, Lau, & Howard, 2015; Nosek, 2012). However, other suggestions have been met with less enthusiasm, including a recent article by Benjamin et al. (2017) advocating that researchers should begin thinking only of p-values less than .005 as "statistically significant", thus changing alpha levels to control Type I error rates. Additionally, P&P (2016) promote the use of fluctuating alpha levels as a function of sample size to assist with these errors. We argue it is not the p-value that needs to be rethought when seeking evidence, but rather what that p-value can tell you in relation to other indicators. While NHST and p-values may have merit, researchers have a wealth of other statistical tools available to them. We believe that improvements may be made to the sciences as a whole when individuals become aware of the tools available to them and how these methods may be used in combination to strengthen understanding and conclusions.

Herein, we have chosen three methodologies to focus on: NHST, Bayes Factor comparisons, and Observation Oriented Modeling. We hope that by discussing these methodologies in terms of a simple statistical analysis researchers will be able to easily compare and contrast methodologies. For this discussion, it is important to understand their historical background, procedural steps, and limitations, which are outlined below. After this discussion, we describe a simulation study comparing methodologies and alpha criteria, and end with a potential implication for researchers.

Null Hypothesis Significance Testing

History

Many attribute the frequentist NHST procedure to Ronald A. Fisher (Fisher, 1932). However, Fisher's ideas are a far cry from the NHST procedure implemented today. Fisher believed in creating one "null" hypothesis, which he described as a hypothesis to be "nullified", or proven incorrect, not as a zero-difference hypothesis (Lehmann, 2011). He also believed that the use of any omnibus level of significance showed a "lack of statistical thinking" (Gigerenzer, Krass, & Vitouch, 2004). He instead believed we should report the exact significance value of a test and let others make their own decision about the claims, which is more in line with the current reporting recommendations provided by the American Psychological Association (APA, 2010). Fisher spoke of this work to William Gosset, the man who created the Student's t-test and contributed work on the correlation coefficient (Lehmann, 2011). Gosset in turn discussed the idea of an alternative hypothesis, a piece not included in Fisher's procedure, with decision theorist Egon Pearson.

From this discussion, Egon Pearson and Jerzy Neyman created Neyman-Pearson decision theory. This theory consists of two hypotheses (i.e., null and alternative) and a binary decision criteria (i.e., significant or not; Lehmann, 1993). However, this procedure created the possibility of researcher decision errors (Dienes, 2008). A researcher may falsely reject the null hypothesis (Type I error, alpha) or falsely fail to reject the null (Type II error, beta). Alpha levels set the binary decision criteria, which are used as the critical p-value for hypothesis testing (i.e., *p* < .05), and are thus seen as evidence to reject the null hypothesis. Beta and power are inherently linked, as the likelihood of finding a true effect increases when beta decreases (Maxwell & Delaney, 2004). Although alpha values can be chosen to be quite small, and methods can decrease beta values as well, a researcher can never know if they have made the correct decision, or a decision error. Thus, Neyman and Pearson clearly state that a hypothesis should not be blindly supported based solely on the estimates of one statistical test, and that replication and reproduction of results are imperative. The recent work of the Open Science Collaboration (2015) has also highlighted the need for replication studies and interpretation of results in an appropriate context. Additionally, Neyman and Pearson emphasized that use of set alphas and betas is illogical and sought instead for researchers to adjust their analysis to the needs of the particular task at hand (Gigerenzer, 2004).

Current NHST Procedure

Neither Fisher's hypothesis testing, nor Neyman-Pearson decision theory quite match the NSHT procedure as it is taught and applied today. Psychologists have largely adopted an amalgamation of the two approaches. Here, we attempt to outline what we believe is the most appropriate way to carry out the traditional NHST procedure, although we note that this is not necessarily how researchers carry out the procedure in practice:

1) Create two hypotheses, one to be “nullified” and one “alternative” hypothesis. Ideally these should be hypotheses that allow for a meaningful conclusion to be reached regardless of whether the null hypothesis is rejected or not.

2) Select an alpha level that is appropriate given the context of your research, your analysis plan, and you research question, and do not blindly adopt an omnibus p-value.

3) Compute your given analysis. If your p-value is less than the chosen alpha, reject the null hypothesis and state that there appear to be differences between your means; however, if your p-value is greater than or equal to the value selected, do not reject the null hypothesis, and state that a difference between the means could not be supported.

Given the mathematical constraints associated with the current NHST procedure as delineated above, there are a number of assumptions that must be met before an analysis is begun (Tabachnick & Fidell, 2007). Data need to have no missing values and no outlying or influential observations. Data must have a normal sampling distribution, be linearly related, and have independent errors. Depending on the statistical test, data must also be checked for equal variances, sphericity, and additivity. These assumptions can be checked and, if necessary, corrected for; however, violations of these assumptions can lead to inaccurate decisions and attenuated power.

While this approach is widely used, there are many limitations associated with it. First, this method can be sensitive to violations of the stated assumptions if the sample size is not large enough to create a normal sampling distribution. These tests are not appropriate for phenomena with non-normal sampling distributions, phenomena that are not linearly related, or those that violate any of the other assumptions mentioned above (Tabachnick & Fidell, 2007). Even if assumptions are met, or nonparametric tests are implemented, this methodology does not allow a researcher to state anything about the absence of an effect (i.e., no true differences). Through NHST, one can only discuss evidence regarding the alternative hypothesis; one can never support the null hypothesis through this procedure (Tabachnick & Fidell, 2007). Given the recent findings regarding reproducibility, showing support for the absence of an effect is even more crucial.

Bayes Factors

History

Thomas Bayes was a statistician and Presbyterian minister whose works are still influential today. Bayes' theorem solved the inverse probability problem, namely that through the frequentist approach, one can only know the probability of data existing given a hypothesis being true, never the probability of a hypothesis being true given that the data exist. Bayes' theorem allows one to calculate the probability of a hypothesis given some data (posterior belief) by using how probable one believes the hypothesis to be before data was collected (prior belief) and how probable one believes the data to be given one's hypothesis (likelihood). Thus, with his theorem, researchers are able to update (through the use of the likelihood) our initial beliefs (our prior) given some data. Pierre-Simon Laplace pioneered Bayesianism and advocated for a broader interpretation of this theorem. The use of Bayesian statistics has been suggested as an NHST alternative, but this approach has largely been undervalued in favor of frequentist methods as, until recently, Bayesian analysis required considerable computational effort. However, today we possess the technology necessary to conduct Bayesian analyses efficiently. While open source software, such as *R* and JASP, require minimal learning to be able to effectively operated (Mourey & Rouder, 2015), researchers will need to invest more effort to understanding the focus and interpretation of Bayes Factor comparisons as they differ from traditional NHST.

The Bayesian framework can be viewed as a continuum, with objective Bayesian analyses on one end, and subjective Bayesian analyses on the other. While this topic could lend itself to its own manuscript, here we will simply summarize the two endpoints, and discuss where our analysis may be perceived to fall on the line. Objective Bayesian analysis is closest to frequentist theory, as priors are set to be as uninformative as possible to allow little, if any, influence on the estimates and distribution of the posterior; thus, the data is allowed to maximally effect the posterior distribution. On the other end, subjective Bayes analyses include rigorously informed priors so that current knowledge can play a large role in the posterior. Our current analysis splits these two; we do not utilize completely uniformed (objective) priors, as we can adjust for basic knowledge of the constraints of our data type. Given the usual lack of information about underlying distributions, a wider band of inclusion was used for prior information. The Bayes Factor package (CITE) assists greatly in the choice of prior and is especially user-friendly for applied researchers, as it makes use of recommended default priors that have been chosen to be safe to assume under a broad range of data and topics (Rouder et al. 2012, Rouder et al., 2009). Instead of conventional F, t, and p-values, a ratio of the likelihood of the alternative model to the null is report, usually BF10. For instance, BF10 = 20 would indicate that the effects model is favored 20 to 1 over the null model. Conversely, if the BF10 were 0.10, the null model is favored 10 to 1 over the effects model.

Current Procedure

The procedure behind Bayes Factor (BF) comparisons requires two steps.

1) One must design two models for the data. For our purposes, the first of these models will be the null model, which states that there are no differences between means. The second model for these analyses is the effects model, which states that each mean is allowed to be different from the grand mean. In designing these models, one must choose the prior distributions that are believed to describe the data. Reasonable expectancies of where the data lie should be incorporated in this decision based on previous research into the studied phenomena (Rouder et al., 2012).

2) Analyze the data given the selected priors and models. Consider the BF and use the BF10 as evidence of how one should update their beliefs about the models.

Based on the flexibility of the analysis, the only assumption that needs to be made is that data exists such that two competing, plausible models with different constraints may be specified. However, given the analysis method we have adopted, we additionally needed to ensure no missing data occurred in our dataset.

Bayesian inference improves upon the traditional frequentist point of view by allowing not only a clear interpretation of the evidence provided by the data, but also the ability to speak in favor of the null hypothesis. However, some limitations do arise in this paradigm. Bayesian analyses require the researcher to take an active role in the choice of prior distributions for the phenomenon they are modeling, and this decision can take some effort to fully understand; however, in the meantime there are packages such as Bayes Factor that allow the researcher simple default options that can readily lend themselves to many research areas with little fear of being outrageous specifications. Further, unlike NHST, Bayesian analyses do not necessarily control long-run error rates, as the focus is on updating current model beliefs. Another concern that many researchers have is that these analyses are necessarily sensitive to prior choice. However, research has shown that the choice of priors has essentially no effect on conclusions when sufficient data has been collected as the priors give way to the weight of the data (Klugkist & Hoijting, 2007; Rouder et al. 2012) and when reasonable priors are considered, data are only mildly sensitive to these (Haaf & Rouder, submitted). Finally, many believe Bayesian analysis to be too computationally intensive to complete. However, many simple programs, packages, and tutorials exist to help ease the transition from frequentist to Bayesian analysis (JASP Team, 2017; Kruschke, 2014; Morey & Rouder, 2015).

Observation Oriented Modeling

History

James Grice argues that our problems as a science go beyond use of NHST and extend into the philosophical ideas underpinning our research. Therefore, he developed a new paradigm called Observation Oriented Modeling (OOM). He reasons that by viewing psychology through the lens of realism, instead of positivism, we should be able to properly and effectively conduct research and analyze data. In contrast to positivism (i.e., which is solely concerned with finding an effect, not with how the effect occurred), realism is the belief that effects conform to their cause and that all theories have an underlying truth. By viewing science as knowing nature through its causes, we can use Aristotle's four causes (material, efficient, formal, and final) to think in terms of forming structures and processes for phenomena. Switching to this philosophy allows for techniques that match the daily activities of social scientists in their endeavors to unravel the story of how humans operate. Using OOM, a researcher does not focus on population parameters and their various underlying assumptions; instead, the researcher is encouraged to take a step back and to focus on observations at the level of the individual.

Generally speaking, this approach can handle any type of data, including ordinal rankings and frequency counts, as all analyses are calculated in the same general fashion. This simplicity occurs because OOM works on the deep structure of the data. Briefly, this means that through observational definition, the program separates these units into binary code. Deep structures can be arranged to form a matrix, which can then be manipulated via matrix algebra, binary Procrustes rotation, and other operations to investigate the data. The most important values from any OOM analysis are the PCC (percent complete match) values. These values represent how well the observations matched the stated or expected pattern or, in the case of causal modeling, how many of the observations conformed to a given cause. Complete matches are the proportion of observations that match the researcher-designated pattern on all dimensions. The PCC value replaces all of the conventional values for effect size used in statistical analyses. The analysis we focus on here (Ordinal Pattern Analysis; OPA) does not form any type of linear or nonlinear equation or regression, but simply looks for those individuals who match the expected ordinal pattern.

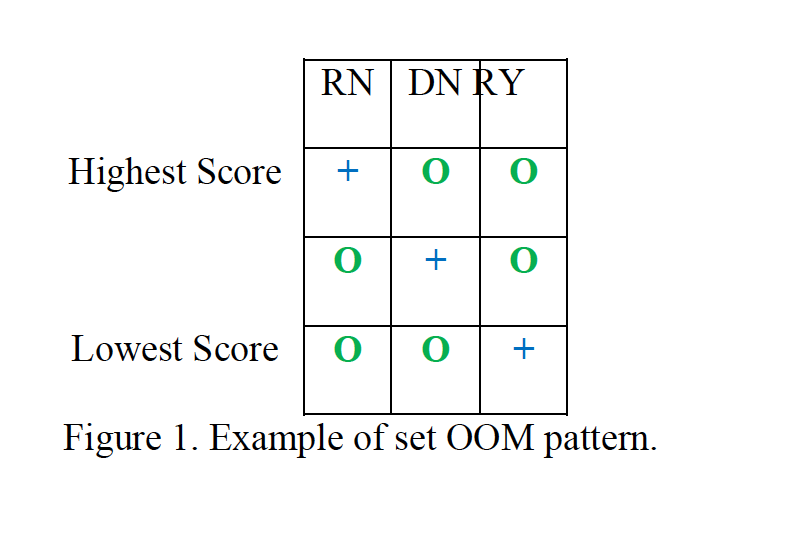
In OOM, p-values are no longer utilized. As a secondary form of reference value, a chance value or c-value, is obtained by randomizing observations a researcher set number of times, often thousands of times. This randomization procedure is akin to permutation tests, where the original data is shuffled a number of times to create a number of comparable data sets. These randomized data sets are then compared to the designated pattern. If the randomized data sets fit the pattern as well as or better than the actual data does, the c-value will be high (close to 1). Low c-values (close to 0) are indicative of distinct observations that are not likely due to chance. Although low c-values, like low p-values, are desirable, c-values do not adhere to a strict cut-off and should be considered a secondary form of confirmation for the researcher that their results are distinct.

Current Procedure

The OPA is analogous to repeated measures ANOVA and contains two steps.

1) Designate the expected ranked pattern: each variable as being higher, lower, or equal to the other variables. See Figure X for an example of a defined pattern (Note that "+"s represent hypothesized squares for the given pattern and "O"s represent non-hypothesized squares).

2) Analyze the data using the OPA. Consider the PCC and c-values in light of the data and use your best judgment as to whether or not the data conform to the expected pattern. This analysis only requires the assumption that the data exists such that a pattern may be designed.



|  |  |  |  |
| --- | --- | --- | --- |
|  | Level 1 | Level 2 | Level 3 |
| Highest Score | O | O | + |
|  | O | + | O |
| Lowest Score | + | O | O |

As with all of these methodologies, limitations do exist. This approach is largely concerned with patterns of responses, not with magnitudes of differences, which may be an integral piece of information to some researchers. Unlike all approaches mentioned before, we do not discuss the probability of some data given our hypothesis here, and instead focus on the observed responses of the individual and how it may or may not behave as expected. Finally, similar to the Bayesian analysis, long-run error rates are not discussed in this methodology.

A Simulation Study

Simulated Data

In this study, we generated 20,000 datasets by manipulating sample size and effect size for a repeated measures design with three levels. These datasets were created using the *mvtnorm* package in *R* (Genz et al., 2017), and all code for simulations can be found at https://osf.io/u9hf4/. Likert data, ranging from 1 to 7, was created by rounding mvtnorm estimates to whole numbers and truncating any data points outside the appropriate range (i.e. values < 1 were rounded to 1, and values > 7 were rounded to 7). The means for each level were set to 2.5, 3.0, and 3.5, and effect sizes were manipulated by adjusting the standard deviation to create negligible effects (*SD* = 3.39, *d* = 0.10), small effects (*SD* = 3.00, *d* = 0.20), medium effects (*SD* = 0.5, *d* = 0.50), and large effects (*SD* = 0.10, *d* = 0.80) using Cohen (1992)'s traditional guidelines for *d* interpretation. The smallest effect size was set such that Likert style data could still be retained with the smallest possible effect size. Sample size was manipulated at 10, 30, 100, 500, and 1,000 data points. All combinations of the five sample sizes and four effect sizes were created and each dataset was simulated 1,000 times, totaling 20,000 datasets.

The advantage of using *mvtnorm* and set *SDs* for each group was the ability to approximate the assumptions of normality by randomly generating from a multivariate normal distribution, and homogeneity by setting equal *SDs* for each group. In a repeated measures design, the assumption of sphericity was met by setting the correlations between levels in *mvtnorm* to zero. By maintaining the lowest level of relationship between levels, we additionally controlled for power and examined situations of significance given the lowest power scenario. During the data simulation, the standard deviation of the difference scores was examined to maintain differences greater than zero, especially for low *n* simulations.

Analyses Performed

Means, mean differences between levels, and the confidence intervals for each mean can be found in the complete dataset online, https://osf.io/u9hf4/. For each simulation, we also calculated *d* values using the standard deviation of the difference score as the denominator (*dz*; Lakens 2013). The *MOTE* library was used to calculate the non-central confidence interval for each *d* value as well (Cumming, 2014, cite ourselves). This data was mainly used to determine if simulations were meeting expected values overall.

Parametric NHST - Repeated Measures ANOVA

Repeated measures ANOVA using the *ezANOVA()* function in the *ez* library was utilized with type three sum of squares (Lawrence, 2017). This style of ANOVA is used to compare the same individuals across multiple or all conditions in an experiment. The null hypothesis states that there are no significant differences between groups, and the research hypothesis posits that there are differences, but does not specify where using the *F* distribution focusing on *p* values.

To determine where differences may exist, *post hoc* dependent *t*-tests are normally analyzed in the event of a significant *F*-ratio. We did not run all pairwise comparisons, instead focusing on the linear trend simulated by comparing level one to two and level two to three. This set of comparisons also controlled the effect size between comparisons, as comparing level one to three would have doubled the effect size. However, we assumed that researchers might compare all three pairwise combinations in practice, and so used a Bonferroni correction across all three possible pairwise combinations to calculate *p* values for *post hoc* tests. Interested readers can find all three comparison values in the complete dataset online. A *p*-value of less than .05 was binned as significant, whereas *p*-values ranging from .10 to .05 were binned as marginally significant. Any *p*-values larger than .10 were binned as non-significant. A second set of *p*-value comparisons was calculated given Benjamin et al. (2017) suggestion to change criterion to less than .005. Any *p*-value less than .005 was binned as significant, while data ranging from .005 to .10 was marginal or suggestive, and *p* > .10 was non-significant.

**Bayesian Analysis: Bayes Factor Comparisons**

We compared a null model with one grand mean for all three levels to an effects model wherein means were allowed to differ using the BayesFactor package (CITE). The default in this package is a Jeffreys prior with a fixed rscale (0.5) and random rscale (1.0). BF were calculated, and follow up *t*-test BFs were computed for the same two comparisons as in the previous models using default priors from the BayesFactor package (e.g., Jeffreys prior for population variance, Cauchy prior for standardized effect size). To compare Bayesian results to other statistical methods, we used recommendations from KASS\_RAFTERY95 to bin results into weak evidence (BFs < 3), positive evidence (e.g., akin to marginal *p*-values, BFs = 3-20), and strong evidence (BFs > 20). BF interpretation should focus on understanding the odds of model ratios, and these bins are used here as a convenient comparison to procedures that do have set criteria for interpretation (cite mourey blog here).

OOM: Ordinal Pattern Analysis

An *R* script of the Ordinal Pattern Analysis from GRICE's OOM program was provided from THE DUDES CITE. We set the expected ranked pattern as level one less than level two less than level three (see Figure X). Once this pattern was defined, we analyzed the data to see if each individual's set of observations matched the set ordinal pattern. PCC values were generated and *c*-values were computed by randomizing the data 1,000 times. Solely for purposes of comparison, we used the following “significance” coding schema: “significant” studies had a high PCC value (.50 < PCC < 1.00) and a low *c*-value (*c* < .05), marginal studies had a high PCC value and a moderate c-value (.05 < *c* < .10), and non-significant studies had low PCC values (PCC < .50), regardless of their *c*-values.

Results

Percent of Estimates

For all simulations, we binned the estimates into significant, marginal, and non-significant effects and calculated the percent of each category of estimates by statistical analysis, sample size, and effect size. These estimates were binned across both the overall and follow up *post hoc* tests, and the combined data is presented for this analysis. Since all three categories of binning total to 100%, we present only the significant and non-significant results. All analyses and findings can be found online at https://osf.io/u9hf4/.

Significant omnibus estimates are presented in Figure 1. For the negligible effects condition at < .05 (solid lines), we found that NHST analyses showed a predictable Type I error bias, in that they detect significant estimates with extremely small *d* values as sample size increases. Binned BF values show a similar pattern, but are more conservative with a smaller percentage of significant estimates. OOM analyses are the most conservative, essentially never detecting an estimate in the no effect simulations. Small effect sizes show the same pattern for NHST, BF, and OOM results, with the proportion of significant estimates increasing more rapidly and asymptoting at a smaller sample size than no effect. At medium effect sizes, NHST analyses nearly always detect differences in estimates, while BF and OOM analyses were considered "significant" around 75% of the time. Interestingly, with large effect sizes, OOM analyses mirror NHST by always detecting estimates, and BF analyses are generally more conservative except at the largest sample size. Figure 1's dashed lines indicate the results if values are binned at *p* < .005, and the differences between these results is very subtle. Lowering reduces the number of significant estimates at small *n* values for all four effect sizes, with a more pronounced difference at negligible and small effect sizes. However, the graphs converge to the same conclusion: large enough sample sizes can produce significant results at negligible and small effect sizes.

Figure 2 portrays the results for non-significant binned simulations, which are the same for both criterion. Across all effect sizes, BF and NHST showed similar results, where non-significant estimates are detected at lower sample sizes for negligible and small effect size simuluations. At medium and large effect sizes, almost all estimates would have been considered significant, therefore, detection rates for non-significant estimates are around zero. OOM displayed a conservative set of findings, showing nearly 100% non-significant estimates at none and small effect sizes (mirroring results from Figure 1). At medium effect sizes, approximately a quarter of estimates were non-significant, illustrating the conservative nature of OOM interpretations.

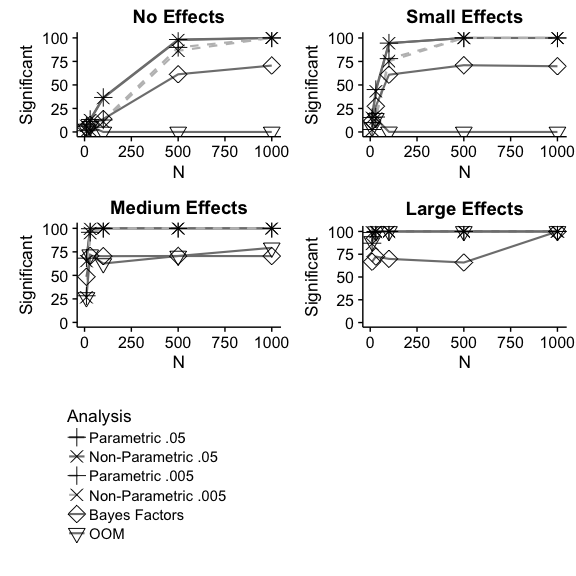


Figure 1 Percent of significant estimates at < .05 (solid) and < .005 (dashed) for each analysis given effect size and sample size.

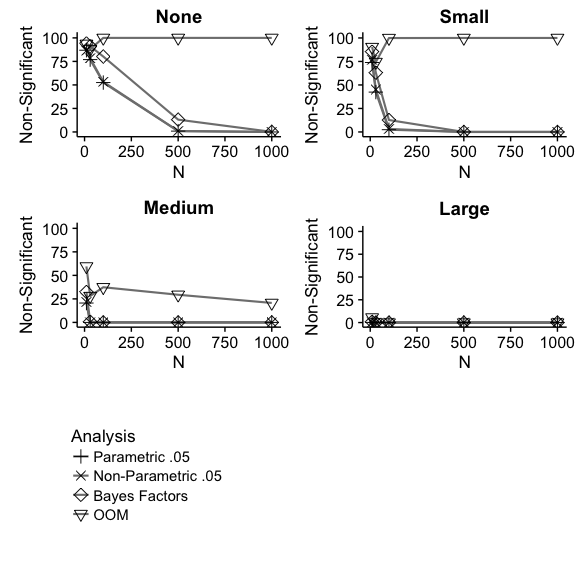


Figure 2 Percent of non-significant effects for each analysis given effect size and sample size.

Percent Agreement

One goal of this project was to expand the toolbox of statistical options for researchers to determine what evidence supports their hypotheses by examining multiple methodologies. We calculated the percent of time that all three analyses agreed across overall and *post hoc* comparison estimates. Figure 3 illustrates the pattern of 100% agreement on effects for omnibus tests only at each sample size, and effect size. Figure 4 portrays the results for *post hoc* tests, which only uses NHST and BF analyses, as OOM does not have a *post hoc* test (i.e., the test is a pattern analysis that presupposes the expected direction of *post hoc* tests).

When effect size was negligible and for small effects, agreement was highest across small samples and decreases across sample size, as NHST was overly biased to report significant estimates and OOM and BF were less likely to do so. For medium and large effect sizes, 50-75% agreement was found, usually regardless of sample size. Additionally, we found that for negligible, small, and medium effects, agreement for *post hoc* tests was higher than agreement for overall comparisons. The *post hoc* comparisons for levels 1 to 2 and levels 2 to 3 were less likely to be binned as significant across none and small effects, so the agreement levels were higher for these individual comparisons due to non-significant follow up tests. The omnibus test was more likely to be significant due to the inclusion of the effect of comparisons between level 1 and 3, which are double the effect size. However, these *post hoc* comparisons do not include the conservative significant binning from OOM, which decreases the omnibus 100% agreement seen in Figure 3. Again, the differences between *p* < .05 and *p* < .005 are minimal. Complete tables of percentages of binning across omnibus and *post hoc* tests, along with agreement percentages broken down by bins can be found at https://osf.io/u9hf4/.

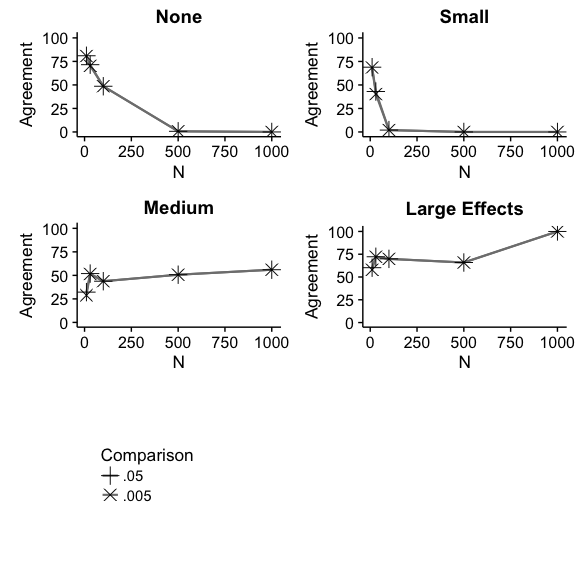


Figure 3 Percent of agreement across each analysis given effect size and sample size for omnnibus tests.

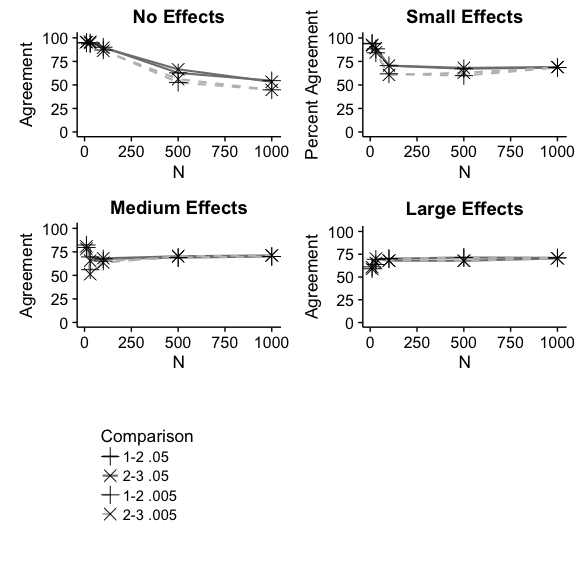


Figure 4 Percent of agreement across each analysis given effect size and sample size tests with < .05 (solid) and < .005 (dashed).

Discussion

This manuscript was designed to showcase available methodologies to researchers and to compare the conclusions each methodology might make in a given data environment. We believe that the application of multiple methodologies might assist in strengthening our conclusions and improving reproducibility by giving researchers the ability to weight various forms of evidence. We found that changing the threshold at which p-values are deemed "significant" had little to no effect on conclusions, especially at large sample sizes, regardless of effect size. This is notable as the article by Benjamin et al. (2017) states that an increase in sample size is likely to decrease false positives “by factors greater than two”, and work by P&P (2016) state that an adaptive level of significance would be beneficial in these circumstances, neither of which are not supported by our simulations. Our science will not grow by moving the significance line in the sand, as this line has already been shown to have "no ontological basis" (R&R1989). Instead, we need to embrace the multitude of perspectives available to us and to begin to use a combination of approaches to qualify the strength of evidence. By comparing multiple methodologies, we can see a more nuanced version of our data. Regardless if analyses agree or disagree on the presence of an effect, a researcher can investigate the size of the effect and discuss conclusions accordingly. Each methodology behaves slightly differently in given data environments, which might begin to highlight meaningful differences when discussed together.

Some may contest that all of these analyses are capable of being hacked, like *p-*values, through researcher degrees of freedom, choice of priors, or pattern choice, among other actions. Transparency throughout the research process is key to eliminating these issues, as alpha changes may only encourage bad research practices with the current incentive structure on publishing. With the Internet, we can share research across the globe, but research often still occurs behind closed doors. The Open Science Framework grants insight into research processes, allowing researchers to share their methodologies, code, design, and other important components of their projects. In addition to posting materials for projects, pre-registration of hypotheses and methodology will be an important facet in scientific accountability. Further, with increased transparency editors and other researchers can weigh the evidence presented according to their own beliefs.

Our key suggestion in this project is the redefinition of evidentiary value. The current focus on *p*-values has shown to be problematic, as many of the studies from the Open Science Collaboration (CITE) do not replicate at *p* < .05 or *p* < .005 (figure out how to cite this). With the change in transparency mentioned above, publishing research with solid research designs and statistics, regardless of *p*-values, will allow for a broader range of evidence to become available.  Publishing null findings is critical in replication and extension for discovering the limits and settings necessary for phenomena. Registered replications and reports will allow studies to be accepted prior to results being known, thus allowing researchers to focus on experimental design and hypotheses apriori instead of p-values posthoc. Reports should describe multiple indicators of evidence, such as effect sizes, confidence intervals, power analyses, Bayes Factors, and other descriptive statistics.

A misunderstanding of statistical power still plagues psychological sciences (Bakker, CITE), and often, individual research labs may not have the means to adequately power a proposed study. Multilab studies and collaboration with other scientists is fundamental to alleviating these issues, while encouraging interdisciplinary science. Collaboration increases our statistical abilities, as every researcher cannot be expected to be proficient in all methods and analyses, but teams of researchers can be assembled to cover a wider range of statistical skills to provide adequate estimates of evidence in their reports. We understand that there may be resistance to the implementation of multiple methodologies as these new methodologies take time and effort to learn. However, through the use of free programs (JASP, R, OOM, shiny) and tutorials (youtube, coursera, statstools.com), we believe all researchers are capable of learning these analyses. We believe that through the expansion of our analytical knowledge and application of these new methodologies, we can begin to attenuate some of the strain currently placed on psychological science and to increase the strength of evidence in our discipline.

References

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:[10.1037//0033-2909.112.1.155](https://doi.org/10.1037//0033-2909.112.1.155)

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:[10.1177/0956797613504966](https://doi.org/10.1177/0956797613504966)

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2017). mvtnorm: Multivariate normal and t distributions. Retrieved from <http://cran.r-project.org/package=mvtnorm>

Lawrence, M. A. (2017). ez: Easy analysis and visualization of factorial experiments. Retrieved from <http://cran.r-project.org/package=ez>

### Nonparametric NHST - Quade's Test

Two well-known dependent samples nonparametric tests are Friedman's test and Quade's test (CITE THESE DUDES). The Quade test was chosen over Friedman's test because Friedman's test has lower power with only three levels (Conover, 1999). Observations are first ranked within each participant (sometimes called blocks), and ties are handled by using the average rank. After ranking, ANOVA and the \*F\* distribution are used to calculate \*p\* values. Quade's test was implemented using \*quade.test()\* in base \*R\*, and \*post hoc\* comparisons were calculated by using the \*posthoc.quade.test\* function in the \*PMCMR\* library (CITE) with a Bonferroni correction for all possible pairwise combinations. \*p\*-values were binned using the same rules as described above, and exact values can be found online.