Beyond *p*-values: Utilizing Multiple Methods to Evaluate Evidence

Kathrene D. Valentine[1], Erin M. Buchanan[2], John E. Scofield[1], & Marshall T. Beauchamp[3]

[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 T. 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. KDV and EMB decided on the study design. MTB helped in initial data analysis. JES and EMB programmed the R code for simulation, analysis, and graphics. KDV wrote the first draft of the manuscript, which EMB put into R markdown. All authors critiqued and edited the manuscript, and all authors approved the submitted version of the manuscript.

Correspondence concerning this article should be addressed to Kathrene D. Valentine, 210 McAlester Ave, Columbia, MO 65211. E-mail: Katy.valentine3@gmail.com

Abstract

Null hypothesis significance testing is frequently cited as a threat to the validity and reproducibility of the social sciences. While many individuals suggest we should focus on altering the *p*-value at which we deem an effect significant, we believe this suggestion is short-sighted. Alternative procedures (i.e., Bayesian analyses and Observation Oriented Modeling) can be more powerful and meaningful to our discipline. However, these methodologies are less frequently utilized and are rarely discussed in combination with NHST. Herein, we compare the possible interpretations of three analyses (ANOVA, Bayes Factor, and an Ordinal Pattern Analysis) in various data environments using a simulation study. The simulation generated 20000 unique datasets which varied sample size (*N*s of 10, 30, 100, 500, 1000), and effect sizes (*d*s of 0.10, 0.20, 0.05, 0.80). Through this simulation, we find that changing the threshold at which *p*-values are considered significant has little to no effect on conclusions. Further, we find that evaluating multiple estimates as evidence of an effect can allow for a more robust and nuanced report of findings. These findings suggest the need to redefine evidentiary value and reporting practices.

Keywords: null hypothesis testing, p-values, Bayes Factors, Observation Oriented Modeling, evidence

Beyond *p*-values: Utilizing Multiple Methods 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 (Etz & Vandekerckhove, 2016; Lindsay, 2015; Open Science Collaboration, 2015; van Elk et al., 2015). 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, Spies, & Motyl, 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 α levels to control Type I error rates. Additionally, Pericchi and Pereira (2016) promote the use of fluctuating α 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. These sentiments have been shared by the American Statistical Association who recently held a conference focusing on going beyond NHST, expanding their previous stance on *p*-values (Wasserstein & Lazar, 2016).

Therefore, we undertook this project to begin to let researchers see the similarities and differences both within the methodological design, as well as within the interpretations of statistics as meaningful. Herein, we have chosen three methodologies to focus on: NHST, Bayes Factor comparisons, and Observation Oriented Modeling. These three approaches will be compared via simulated data using a repeated measures design with a Likert-type scale as the outcome variable. The aims of this study will be to discuss the conclusions that these three methods would make given the same data, and to compare how often these methodologies agree within different data environments (i.e. given different sample sizes and effect sizes). 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 α criteria, and end with a potential implications 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 shown 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, Krauss, & 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 typical reporting recommendations provided by the American Psychological Association (American Psychological Association, 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, α) or falsely fail to reject the null (Type II error, β). α 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. β and power are inherently linked, as the likelihood of finding a true effect increases when beta decreases (Maxwell & Delaney, 2004). Although α values can be chosen to be quite small, and methods can decrease β 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 αs and βs is illogical and sought instead for researchers to adjust their analysis to the needs of the particular task at hand (Gigerenzer, 2004).

Typical 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. These can be operationalized as follows:

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

3) Compute your given analysis and identify the corresponding p-value. If your p-value is less than the chosen α, 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.

While the NHST procedure itself gives us the testable models, the specific analysis used to test these models here—the repeated measures ANOVA with 3 levels—requires some additional assumptions that must be met before an analysis is begun (Tabachnick & Fidell, 2012). 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, 2012). 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. Given the recent findings regarding reproducibility, showing support for the absence of an effect is even more crucial (Bakker, van Dijk, & Wicherts, 2012; Lakens, 2017).

Bayes Factors

History

Thomas Bayes was a statistician and Presbyterian minister whose works are still influential today (Bellhouse, 2004). 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 (Dienes, 2008). 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 (Gelman, 2004). Pierre-Simon Laplace pioneered Bayesianism and advocated for a broader interpretation of this theorem (De Laplace, 1774). The use of Bayesian statistics has been suggested as an NHST alternative (Dienes, 2014; Wagenmakers, 2007), 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 (Morey & 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 (Press, 2002). 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 *BayesFactor* package (Morey & Rouder, 2015) 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, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009). Instead of conventional F, t, and p-values, a ratio of the likelihood of the alternative model to the null is report, usually BF_{10}. For instance, $BF_{10} = 20$ would indicate that the effects model is favored 20 to 1 over the null model. Conversely, if the BF_{10} were 0.10, the null model is favored 10 to 1 over the effects model.

Typical 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). These can be operationalized as follows:

2) Analyze the data given the selected priors and models. Consider the BF and use the BF_{10} 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. While not an assumption of the method, we did additionally needed to ensure no missing data occurred in our dataset as this was a requirement of the package utilized in the simulations.

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. It is important to note that while previous work has indicated that p-values and BF largely agree on which hypothesis should be supported, they differ in the strength of that conclusion, especially when p-values were slightly lower than α (i.e., .05 to .01; Wetzels et al., 2011). 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 *BayesFactor* 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 & Hoijtink, 2007; Rouder et al., 2012) and when reasonable priors are considered, data are only mildly sensitive to these (Haaf & Rouder, 2017). 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, James W. Grice, 2011, 2014; James W. Grice, Barrett, Schlimgen, & Abramson, 2012). 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 given the correct models of these processes we can begin to understand our reality. 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 structures and processes in order to explain 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 the various assumptions underlying statistical tests (e.g., random sampling, normality, homogeneity of population treatment differences, etc.). Instead, the researcher alternatively focuses 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 (see K. D. Valentine & Buchanan, 2013 for an example). This simplicity occurs because OOM works on the deep structure of the data. 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 correct classification) values. These values represent the summation of how well an individual's responses matched the stated or expected pattern or, in the case of causal modeling, how many of the individual's conformed to a given cause. Complete matches are the proportion of observations that match the researcher-designated pattern on all dimensions. For example, in a three-level Ordinal Pattern Analysis (OPA), a person would be tallied as a "complete match" if the ordinal pattern of his/her data matched the expected ordinal pattern across all three levels. Imagine we have set a pattern that designates that time 1 responses should be less than time 2 which should be less than time 3. Given the data for two hypothetical individuals in Table 1, we can see that person a matched the pattern completely, and therefore would be counted in the PCC value. However, while person b matched the first part of our pattern (time 1 less than time 2) they did not match on the third point of our pattern (time 2 less than time 3); thus, they would not be counted in the PCC value. The *PCC* value replaces all of the conventional values for effect size used in statistical analyses. The analysis we focus on here (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 (J. W. Grice, Craig, & Abramson, 2015).

In OOM, *p*-values are no longer utilized (James W. Grice, 2011). As a secondary form of reference value, a chance value (*c*-value) is obtained, which is a type of randomization test in which the researcher determines the number of randomized trials for the test (e.g. 1000 or 5000 randomized versions of actual observations). This procedure is akin to permutation tests, where the original data is shuffled a number of times to create 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) indicate a pattern of observations that is improbable (i.e., unlikely produced by chance) when compared to randomized versions of the same data. 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.

Typical 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 1 for an example of a defined pattern.

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.

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. A repeated measures design was chosen as it is widely used across many disciplines of psychology. 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/. Interested readers can easily adapt the *R* code to incorporate different research designs. Likert data, ranging from 1 to 7, was created by rounding *mvtnorm* estimates to whole numbers and truncating any data points out the appropriate range (i.e. values < 1 were rounded to 1, and values > 7 were rounded to 7). The population 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 ($SD = 0.50$, $d = 0.50$), 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

Descriptive Statistics. 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 (d_z, Lakens, 2013). The *MOTE* library was used to calculate the non-central confidence interval for each d value as well (E. M. Buchanan, Valentine, & Scofield, 2017; Cumming, 2014). 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 population means, and the research hypothesis posts that there are differences between population means, but does not specify which population means may differ, just that one or more will differ as the alternative. This uses 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 the typical researchers might compare all three pairwise combinations in practice and used a Bonferroni correction across all three possible pairwise combinations to calculate p values for *post hoc* tests. Therefore, while we only discuss the two comparisons, we utilized the more stringent cutoff of the Bonferroni correction as we believe this is how the majority of researchers would handle the data. 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)'s suggestion to change α 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. 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 (Morey & Rouder, 2015). 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 and Raftery (1995) 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 (Morey, 2015).

OOM: Ordinal Pattern Analysis. An *R* script of the Ordinal Pattern Analysis from J. W. Grice et al. (2015)'s OOM program was provided from Sauer and Luebke (2016). We set the expected ranked pattern as level one less than level two less than level three (see Figure @ref:(fig:oom-pic). Once this pattern is defined, the we analyzed the data to see if each individual's set of observations match this expected 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 first binned the estimates into significant, marginal, and non-significant effect categories as described in the Analyses Performed section above. Next, we calculated the percentage of these analyses that would be classified into each of these categories, separated about 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 critical omnibus estimates are presented in Figure 2. For negligible effects at $p < .05$ (solid lines), we found that NSHT 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 less percent 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 negligible effects. At medium effect sizes, NHST analyses nearly always detect estimates, while BF and OOM analyses will be 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 2's dashed lines indicate the results if values are binned at $p < .005$, and the differences between these results is very subtle. Lowering α reduces the number of significant estimates at small n values for all four effect sizes, with a more pronounced differences at no and small effect sizes. However, the graphs converge to the same conclusion that large enough sample sizes can produce significant results at negligible and small effect sizes.

Figure 3 portrays the results for non-significant binned simulations, which are the same for α 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 simulations. 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 2). At medium effect sizes, approximately a quarter of estimates were non-significant, illustrating the conservative nature of OOM interpretations.

Percent Agreement

A goal of this project was to expand the toolbox of options for researchers to determine what evidence supports their hypotheses by examining multiple methodologies. We calculated the percent of time that all analyses agreed across overall and *post hoc* comparison estimates. Figure 4 illustrates the pattern of 100% agreement on effects for critical omnibus tests only at each sample size, and effect size. Figure 5 portrays the results for *post hoc* tests, which only uses NHST and Bayes Factor 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 best 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 negligible and small effects, so the agreement levels were higher for these individual comparisons due to non-significant follow up tests. The critical omnibus test was more likely to be significant due to the inclusion of 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 critical omnibus 100% agreement seen in Figure 4. Again, the differences between $p < .05$ and $p < .005$ are minimal. Complete tables of percentages of binning across critical omnibus and *post hoc* tests, along with agreement percentages broken down by bins can be found at https://osf.io/u9hf4/.

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 finding 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" (p. 10), and work by Pericchi and Pereira (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" (Rosnow & Rosenthal, 1989, p. 1277). 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 (Simmons et al., 2011). Transparency throughout the research process is key to eliminating these issues, as α 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 (2015) do not replicate at $p<.05$ or $p < .005$ (Lakens et al., 2017). 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 *post hoc*. Reports should describe multiple indicators of evidence, such as effect sizes, confidence intervals, power analyses, Bayes Factors, and other descriptive statistics (Finkel, Eastwick, & Reis, 2015; Nosek & Lakens, 2014; Van't Veer & Giner-Sorolla, 2016).

A misunderstanding of statistical power still plagues psychological sciences (Bakker, Hartgerink, Wicherts, & van der Maas, 2016), 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, http://www.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.

Limitations

Within any study a number of limitations exist. The largest limitation of our study is that we chose such a narrow focus. Given that we only focused on one analytical design—repeated measure ANOVA with 3 levels—it is possible that these same relationships may or may not exist in alternative design contexts. Additionally, our choices for classification of “significant” effects for *p*-values, Bayesian factors, PCC, and *c*-values was based on what we believe a reasonable researcher may designate; however these classifications may vary in the real world and thus would necessarily alter the conclusions derived here. Finally, due to the specification of our simulation we did not violate any statistical assumptions. It is possible—and highly likely—that violation of these assumptions may cause disruptions in the relationships we see here.

References

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed., p. 272). American Psychological Association. Retrieved from http://www.apastyle.org/products/4200067.aspx

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

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. doi:10.1177/1745691612459060

Bellhouse, D. R. (2004). The Reverend Thomas Bayes, FRS: A Biography to celebrate the tercentenary of his birth. *Statistical Science*, *19*(1), 3–43. doi:10.1214/088342304000000189

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., & Johnson, V. E. (2017). Redefine statistical significance. *PsyArxiv*, (July 22), 1–18. doi:10.17605/OSF.IO/MKY9J

Buchanan, E. M., Valentine, K. D., & Scofield, J. E. (2017). MOTE. Retrieved from https://github.com/doomlab/MOTE

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

Cumming, G. (2008). Replication and p intervals. *Perspectives on Psychological Science*, *3*(4), 286–300. doi:10.1111/j.1745-6924.2008.00079.x

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

De Laplace, P. S. (1774). Mémoire sur les suites récurro-récurrentes et sur leurs usages dans la théorie des hasards. *Mém. Acad. R. Sci. Paris*, *6*(8), 353–371. Retrieved from http://cerebro.cs.xu.edu/math/Sources/Laplace/recurro{_}recurrentes.pdf

Dienes, Z. (2008). *Understanding psychology as a science: an introduction to scientific and statistical inference*. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. doi:10.3389/fpsyg.2014.00781

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), 1–12. doi:10.1371/journal.pone.0149794

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*(2), 275–297. doi:10.1037/pspi0000007

Fisher, R. A. (1932). Inverse probability and the use of likelihood. *Mathematical Proceedings of the Cambridge Philosophical Society*, *28*(03), 257. doi:10.1017/S0305004100010094

Gelman, A. (2004). *Bayesian data analysis* (p. 668). Chapman & Hall/CRC. Retrieved from https://www.crcpress.com/Bayesian-Data-Analysis-Second-Edition/Gelman-Carlin-Stern-Rubin/p/book/9781584883883

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

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In *The sage handbook of quantitative methodology for the social sciences* (pp. 392–409). Thousand Oaks, CA: SAGE Publications, Inc. doi:10.4135/9781412986311.n21

Grice, J. W. (2011). *Observation oriented modeling : analysis of cause in the behavioral* *sciences* (p. 242). Elsevier/Academic Press.

Grice, J. W. (2014). Observation Oriented Modeling: Preparing students for research in the 21st century. *Comprehensive Psychology*, *3*, 05.08.IT.3.3. doi:10.2466/05.08.IT.3.3

Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences*, *2*(4), 1–22. doi:10.3390/bs2010001

Grice, J. W., Craig, D. P. A., & Abramson, C. I. (2015). A simple and transparent alternative to repeated measures ANOVA. *SAGE Open*, *5*(3), 2158244015604192. doi:10.1177/2158244015604192

Haaf, J., & Rouder, J. N. (2017). *Developing constraint in bayesian mixed models.* doi:10.17605/OSF.IO/KTJNQ

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

JASP Team. (2017). JASP. Retrieved from https://jasp-stats.org/

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.1080/01621459.1995.10476572

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*(12), 6367–6379. doi:10.1016/j.csda.2007.01.024

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00863

Lakens, D. (2017). Equivalence tests. *Social Psychological and Personality Science*, *8*(4), 355–362. doi:10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., …

Zwaan, R. A. (2017). *Justifying, not redefining, alpha.* Retrieved from https://osf.io/by2kc

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

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, *88*(424), 1242–1249. doi:10.1080/01621459.1993.10476404

Lehmann, E. L. (2011). *Fisher, Neyman, and the creation of classical statistics.* New York, NY: Springer.

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

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. doi:10.1037/a0039400

Morey, R. D. (2015). On verbal categories for the interpretation of Bayes factors. Retrieved from http://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes Factors for common designs. Retrieved from https://cran.r-project.org/package=BayesFactor

Nosek, B. A., & Lakens, D. (2014). Registered reports. *Social Psychology*, *45*(3), 137–141. doi:10.1027/1864-9335/a000192

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia. *Perspectives on Psychological Science*, *7*(6), 615–631. doi:10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science.

Science, *349*(6251), aac4716–aac4716. doi:10.1126/science.aac4716

Pericchi, L., & Pereira, C. (2016). Adaptative significance levels using optimal decision rules: Balancing by weighting the error probabilities. *Brazilian Journal of Probability and Statistics*, *30*(1), 70–90. doi:10.1214/14-BJPS257

Press, S. J. (Ed.). (2002). *Subjective and Objective Bayesian Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470317105

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284. doi:10.1037/0003-066X.44.10.1276

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi:10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:10.3758/PBR.16.2.225

Sauer, S., & Luebke, K. (2016, January). Observation Oriented Modeling revised from a statistical point of view. doi:10.17605/OSF.IO/3J4XR

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston, MA: Pearson.

Valentine, K. D., & Buchanan, E. M. (2013). JAM-boree: An application of observation oriented modelling to judgements of associative memory. *Journal of Cognitive Psychology*, *25*(4), 400–422. doi:10.1080/20445911.2013.775120

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

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. doi:10.1016/j.jesp.2016.03.004

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:10.3758/BF03194105

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p -values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. doi:10.1080/00031305.2016.1154108

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology. *Perspectives on Psychological Science*, *6*(3), 291–298. doi:10.1177/1745691611406923

Table 1

OOM Ordinal Pattern Analysis Example

1	2	3	4
Individual	Time 1	Time 2	Time 3
A	3	4	5
B	4	5	2

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

Figure 1. Figure of designed Ordinal Pattern Analysis for our simulation student. +s represent hypothesized squares for the given pattern and Os represent non-hypothesized squares.

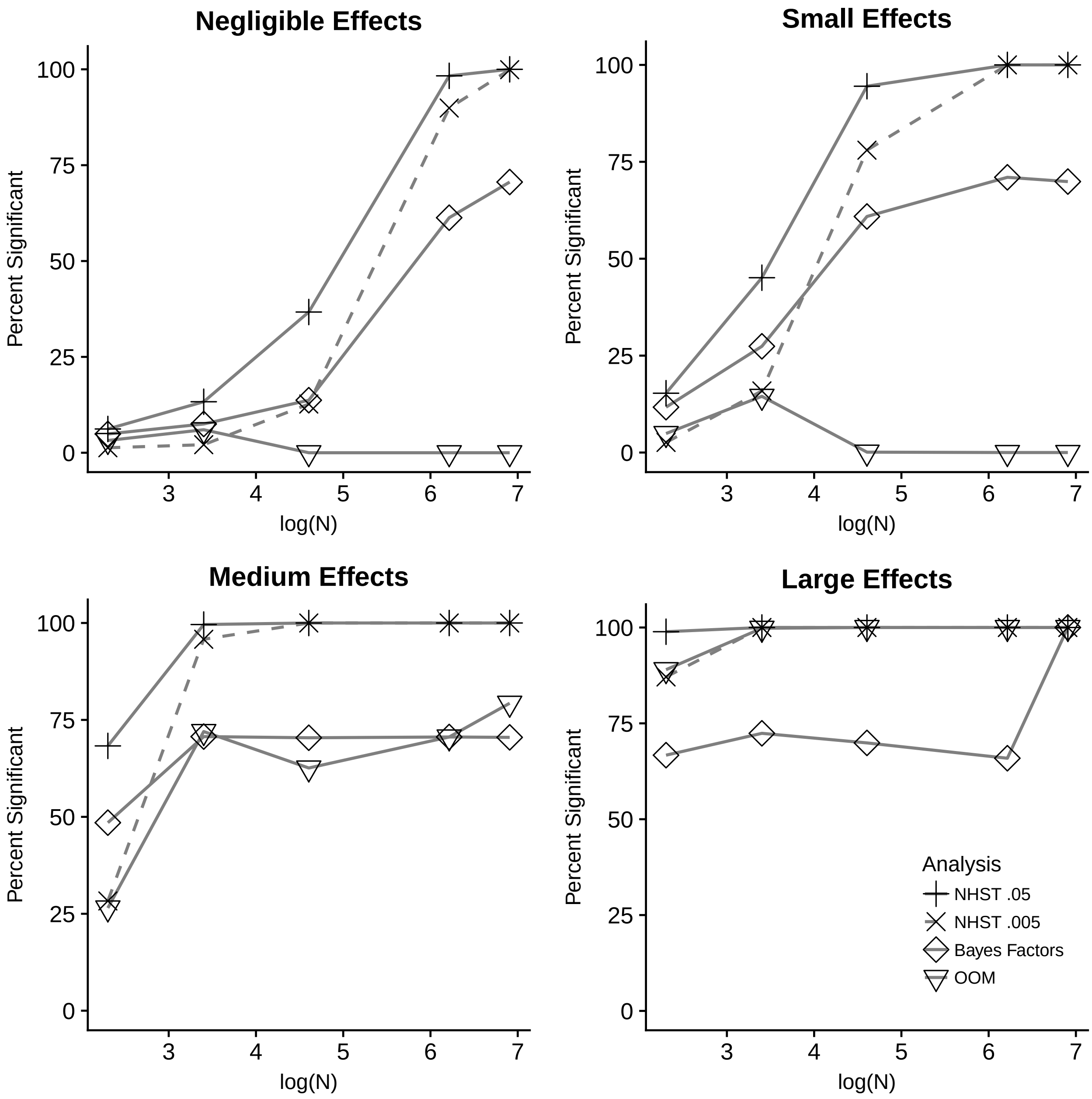

Figure 2. Percent of significant estimates at $p < .05$ (solid) and $p < .005$ (dashed) for each analysis given effect size and sample size.

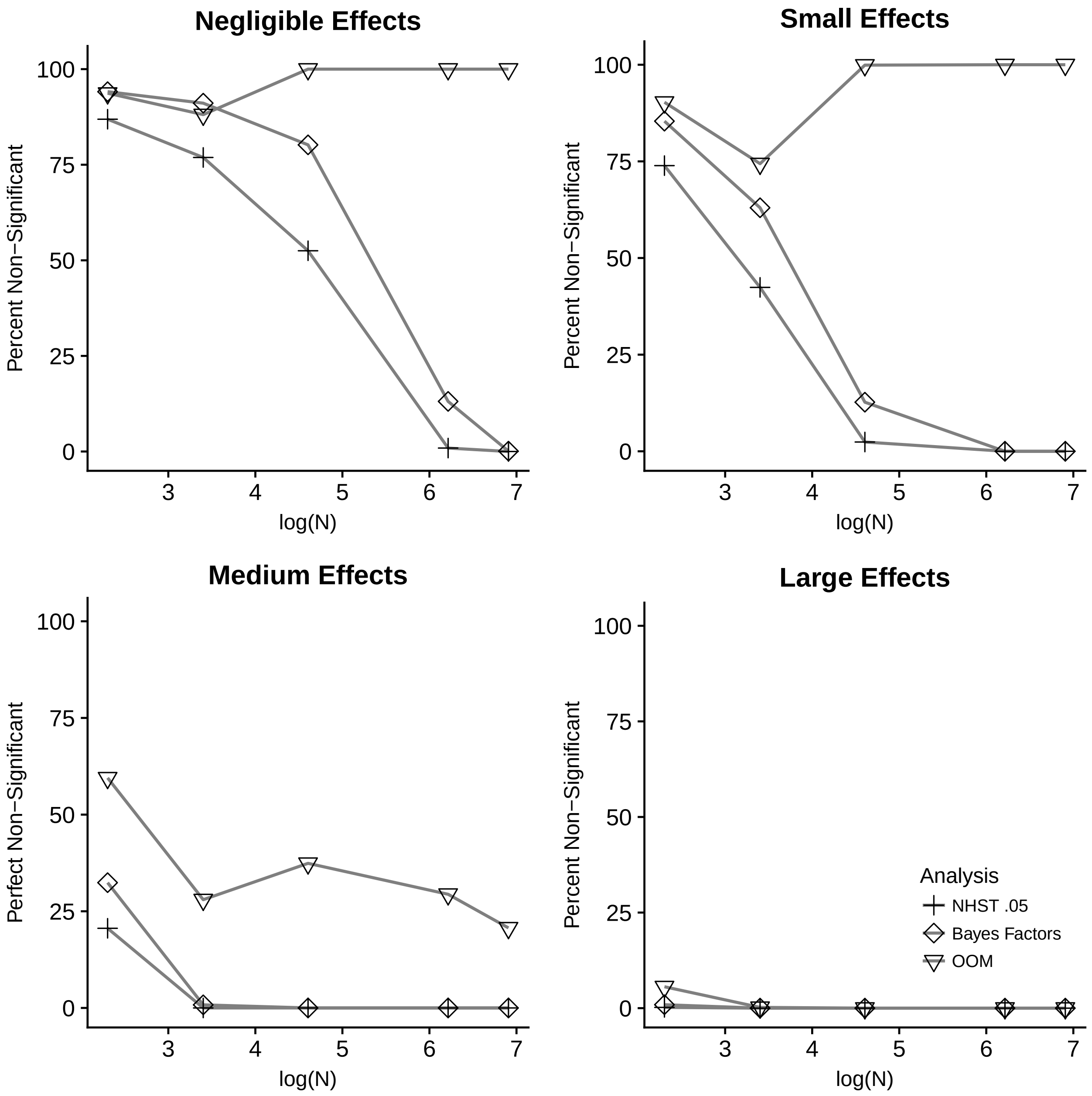

Figure 3. Percent of non-significant effects for each analysis given effect size and sample size.

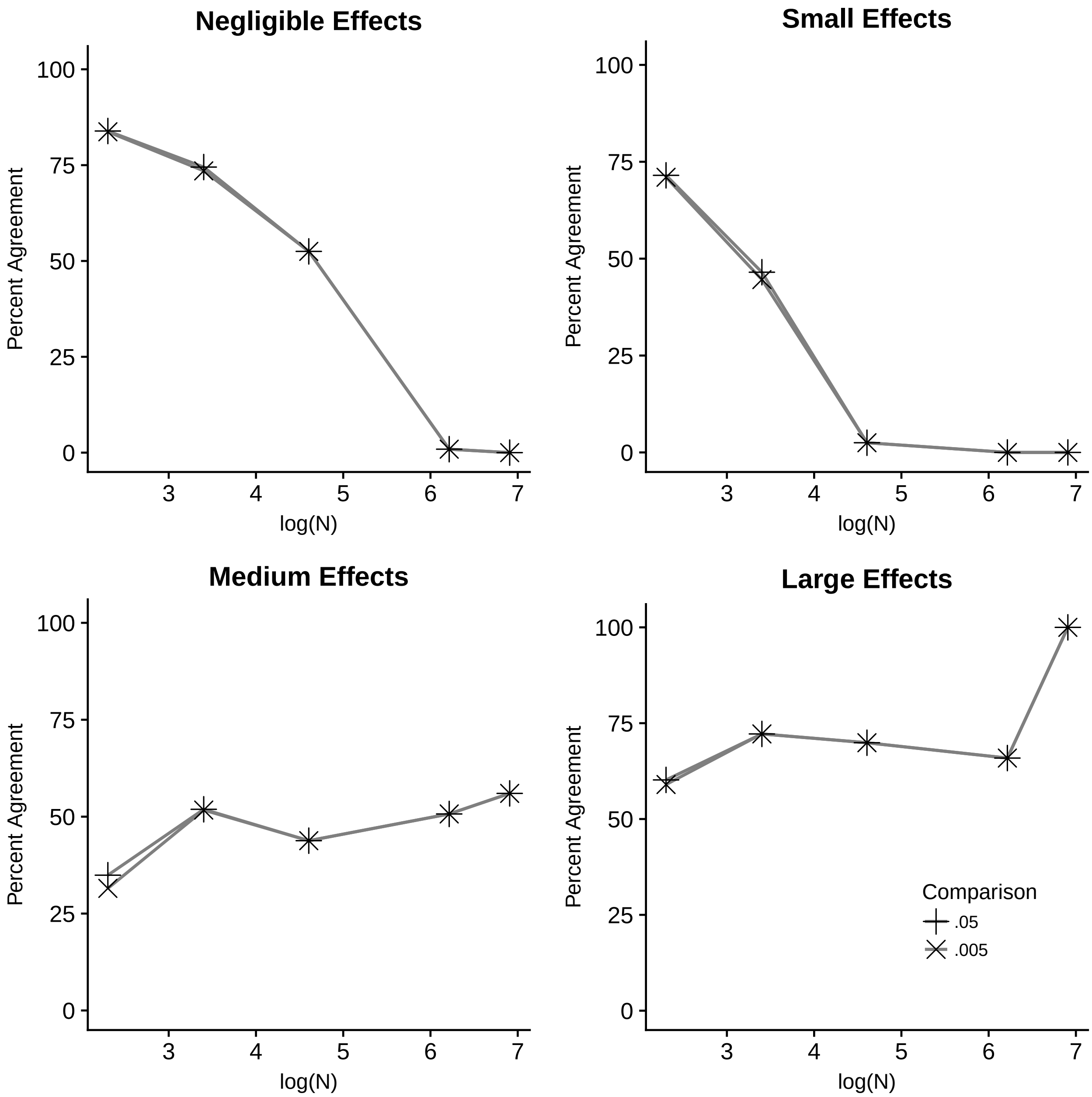

Figure 4. Percent of agreement across each analysis given effect size and sample size for omnnibus tests.

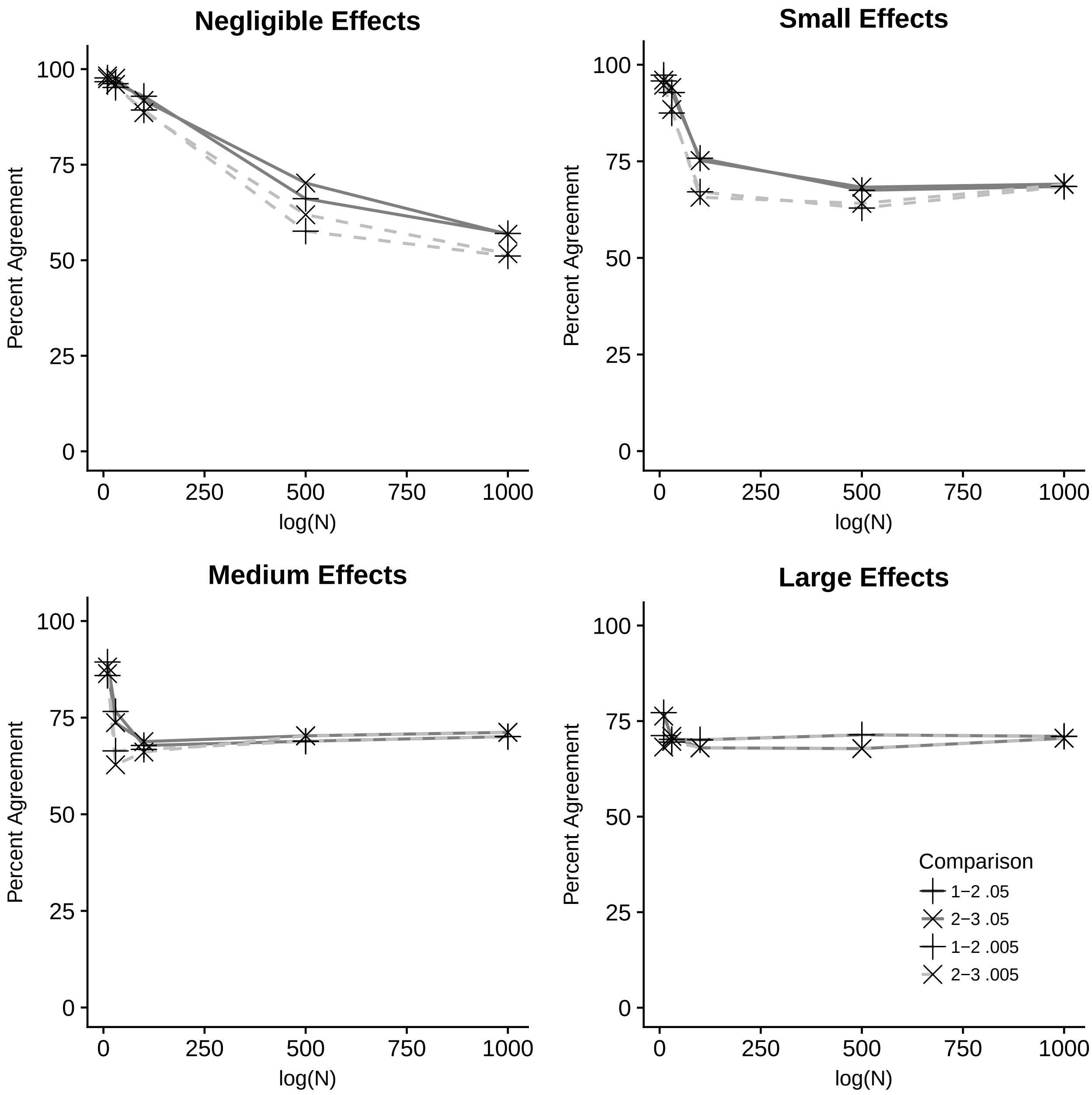

Figure 5. Percent of agreement across each analysis given effect size and sample size *posthoc* tests with $p < .05$ (solid) and $p < .005$ (dashed).