Journal of the American Statistical Association

September 2017 • Volume 112 • Number 519

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Rejoinder: Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane & David Gal

including frequentist and NHST-based inference. As has been intensively discussed elsewhere, we are likely to be increasingly working with extensive volumes of fine-scale data on the systems we study. It has also been noted that "big data needs big models" (Gelman 2014). These big models, including models derived from machine learning methods, as well as flexible procedures deriving from classical statistics such as semiparametric, empirical likelihood, dimension reduction, and localized methods, can be powerful tools for improving the properties of NHST. Recent work on high dimensional inference is providing new tools to build such models while not saturating the models to the point where parameter estimates become meaningless. However, most applied researchers and many statisticians are not using these new tools to their full potential. The findings of McShane and Gal make clear that in terms of communication, training, and methods development, there is still a lot of room to grow.

Funding

The authors gratefully acknowledge funding from the National Science Foundation (DMS-1555141, DMS-1557733, DMS-1513579, DMS-1317631) and the National Institutes of Health (P01 CA142538).

References

Barr, D. J., Levy, R., Scheepers, C., and Tily, Harry, J. (2013), "Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal," *Journal of Memory and Language*, 68. [903]

Candès, E., and Barber, R. (2015), "Controlling the False Discovery Rate via Knockoffs," *Annals of Statistics*, 43, 2055–2085. [903]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [903]

Gelman, A. (2014), "Big Data Needs Big Model," available at *http://andrewgelman.com/2014/05/22/big-data-needs-big-model*. [904]

Howick, J., Friedemann, C., Tsakok, M., Watson, R., Tsakok, T., Thomas, J., Perera, R., Fleming, S., and Heneghan, C. (2013), "Are Treatments More Effective than Placebos? A Systematic Review and Meta-Analysis," *PLoS One*, 11, e0147354. [903]

Knutsson, H., Eklund, A., and Nichols, T. E. (2016), "Cluster Failure: Why fmri Inferences for Spatial Extent Have Inflated False-Positive Rates," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7900–7905. [903]

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A., (2010), "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data," *Nature Reviews Genetics*, 11, 733–739. [903]

Pritchard, J. K., and Voight, B. F. (2005), "Confounding From Cryptic Relatedness in Case–Control Association Studies," *PLoS Genetics*, 1, 302–311. [903]

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, Applications and Case Studies
https://doi.org/10.1080/01621459.2017.1323642

Rejoinder: Statistical Significance and the Dichotomization of Evidence

Blakeley B. McShane[a] and David Gal[b]

[a]Kellogg School of Management, Northwestern University, Evanston, IL; [b]College of Business Administration, University of Illinois at Chicago, Chicago, IL

We heartily thank editor Montserrat Fuentes for selecting our article (McShane and Gal 2017) for discussion. We are grateful for the opportunity to receive feedback on our work from four sets of distinguished discussants who possess a tremendous breadth of knowledge and expertise, and we deeply thank them for the time and effort they put into contemplating and responding to our article. We were delighted that our principal point—namely, that even expert statisticians are sometimes prone to misuse and misinterpret p-values and that these errors disproportionally arise from interpreting evidence dichotomously based on whether or not a p-value crosses the conventional 0.05 threshold for statistical significance—was both clear to and appreciated by our four sets of discussants.

In this rejoinder, we aim to do three things. First, we clarify and expound on certain aspects of our study designs and results to respond to some potential alternative accounts and criticisms raised in the discussion. Second, we tie together several broad themes that emerged in the discussion. Finally, we explore issues related to statistical significance and the dichotomization of evidence in the domain in which we most often work, namely, social psychology and consumer behavior.

In the remainder of this rejoinder, we abbreviate the discussions as DAB (Berry 2017), WMB (Briggs 2017), GC (Gelman and Carlin 2017), and LS (Laber and Shedden 2017).

1. Study Designs and Results

1.1. Study 1

DAB and LS both raise a concern regarding a potential misinterpretation by our subjects of the principal question asked in Study 1, in particular a confusion over whether the question we asked was one about the sample (i.e., about descriptive statistics) or about the population (i.e., about statistical inference). Both key in on the phrase "Speaking only of the subjects who took part in this particular study" used in the response options as

CONTACT Blakeley B. McShane b-mcshane@kellogg.northwestern.edu Associate Professor, Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60611.

potentially responsible for our results, with DAB regarding the phrase as ambiguous and LS regarding it as an "unmistakable cue" (at least *ex post*). We note that, due to the design of our study, this phrase—and its ambiguity or clarity—cannot be responsible for our results. The reason for this is (i) subjects were randomized to one of two wordings of the response options where response wording one included the phrase and response wording two omitted it (this was the sole difference between the two response wordings) and (ii) the results were not substantially affected by the response wording (see Figure 1a of our article; see also Figure 1b, which shows the same was true in Study 1 of McShane and Gal (2016) where a third response wording was used and subjects were authors of articles published in the *New England Journal of Medicine*).

However, it is possible that a different confusion between sample and population may have arisen. In particular, while responses in treatment and control groups are often modeled using infinite population parametric models (e.g., independent normal with different means or independent binomial with different proportions), randomization secures only a finite population permutation model: under randomization, the population in question does not consist of additional subjects who were not included in the study but rather consists of both potential outcomes (i.e., under treatment and under control of which of course only one is observed) of each subject included in the study (generalization to additional subjects is a distinct matter). Under the permutation model, it could be argued that statements such as "the average for the treatment" can be ambiguous in terms of whether they refer to the average for those subjects who actually received the treatment in the study (i.e., the sample average) versus the average for all subjects under the hypothetical that they all received the treatment (i.e., the population average); under the latter interpretation, one might perhaps be justified in giving a different response for the $p = 0.01$ and $p = 0.27$ versions of the question. However, as only four subjects reported using the permutation model, this explanation cannot hold in practice. Further, our response wording generally precluded the latter interpretation (i.e., by asking about the average of "participants who were in Group A" it is unreasonable to assume we were asking about a hypothetical under which all participants were assigned to Group A).

We also wish to reiterate that the claim that the mere presence of a p-value in the question naturally led our subjects to focus on statistical inference rather than description is not really a criticism but rather is essentially our point: our subjects are so trained to focus on statistical significance that the mere presence of a p-value leads them to automatically view everything through the lens of the null hypothesis significance testing (NHST) paradigm—even in cases where it is unwarranted.

Further, as acknowledged by LS, subjects were asked to explain in their own words why they chose the options they chose. As shown in our article, their text responses emphasized that they were thinking dichotomously in a manner consistent with the dichotomization of evidence intrinsic to the NHST paradigm. Moreover, their responses to the two versions of our principal question did not associate particularly strongly with their responses to our various follow-up questions.

A final concern raised by DAB was that our study had a "low response rate" due to potential confusion over whether the question we asked was one about the sample or about the population and generated by the "Speaking only" phrase. In response, we note that our response rate of 27% is actually rather high for this kind of survey and subject population. Further, due to the design of our study, this phrase cannot be responsible for our response rate as (i) subjects were randomized to one of two wordings of the response options where response wording one included the phrase and response wording two omitted it and (ii) those randomized to response wording one would have seen the phrase only after they had already responded to the survey. Instead, if the phrase were to have had an impact, it would have been on the completion rate of our survey rather than the response rate to it. However, our completion rate did not substantially differ by the response wording and, at 94%, is extremely high.

1.2. Study 2

DAB accepts the results of Study 2 for Bayesians but not for frequentists. We do not necessarily disagree with his underlying logic but wish to expound upon this. First, subjects' responses to our follow-up question regarding statistical approach (frequentist, Bayesian, neither, or both) did not particularly strongly associate with their responses to either of the principal questions. Second, the text responses of our subjects provide little support for any concern about Bayesian versus frequentist reasoning. Third, any concern about Bayesian versus frequentist reasoning seems most germane to the likelihood judgment question rather than the choice question. However, as noted in our article and by LS, there is a sense in which option *D* is the correct frequentist option for the likelihood judgment question (i.e., because at no particular p-value is the null hypothesis definitively overturned). More specifically, which drug is "more likely" to result in recovery depends upon the parameters governing the probability of recovery for each drug, and these parameters are unknown and unknowable under a classical frequentist interpretation of the question. However, subjects generally chose option *A* for the likelihood judgment question when the p-value was set below 0.05 but option *D* when it was set above 0.05 rather than option *D* regardless. Thus, it seems improbable that subjects approached the question in this manner.

LS suggest that perhaps some of our subjects were engaging in response substitution (Gal and Rucker 2011), in particular, that subjects who were presented with a p-value greater than 0.05 "'read' between the 'lines' and answer[ed] the question they felt the investigators meant to ask," namely, one of statistical significance. Were subjects engaging in response substitution, we might have expected their text responses to reflect it. In particular, we might have expected them to say something along the lines of, "Drug A is more likely to lead to recovery from the disease than Drug B, but it is not statistically significantly more likely to lead to recovery." However, we did not see text responses of this sort. We further note that, while the likelihood judgment question may allow for this interpretation, the choice question allows little room for it; nonetheless, a meaningful share of subjects did not choose Drug A.

We also note that, while we agree with DAB that "For a different prior distribution it is quite possible for the [posterior] probability that Drug A is more effective than Drug B to be 0.99, say, and yet Drug B have the greater [posterior] mean and so be the correct choice" (although for it to be "correct" requires some additional assumptions about the loss function), we believe this is not relevant to our study as the question wording explicitly encouraged a noninformative prior (i.e., "Assuming no prior studies have been conducted with these drugs, which of the following statements is most accurate?").

2. Themes

There were several broad themes that emerged in the discussion to which we would like to draw attention. There was agreement that the very definition or logic of the *p*-value is problematic in and of itself. This was put perhaps with greatest flourish by DAB who stated that *p*-values are "perversions of logic" that are "fundamentally un-understandable" and led some to the conclusion that "the only reasonable path forward is to kill [*p*-values]" (DAB) and that "there are no good reasons nor good ways to use *p*-values. They should be retired forthwith" (WMB). GC go further and argue that many oft-suggested replacements for *p*-values such as confidence intervals and Bayes factors share some of the same problems in terms of inducing dichotomous (or more broadly categorical) thinking.

Related to dichotomous thinking is what GC term deterministic thinking, namely, "demanding more certainty than [the] data can legitimately supply" (GC) and the related "mentality that $p < 0.05$ means true and $p > 0.05$ means not true" (DAB) (or, as we put it, the assignment of evidence to the different categories "statistically significant" and "not statistically significant" naturally leads to the conclusion that the treatments thusly assigned are categorically different). This becomes particularly problematic and pronounced when, as GC note, most effects measured in applied research represent a mean in some population (or something similar such as a regression coefficient)—a fact which they note "seems to be lost from consciousness when researchers slip into binary statements about there being 'an effect' or 'no effect' as if they are writing about constants of nature;" this issue is strongly compounded in the biomedical and social sciences where an effect (i.e., mean, regression coefficient) of zero is generally implausible.

Hypothesized zero mean effects tie nicely to the issue of the "straw man" null hypotheses decried by LS (but used in the overwhelming majority of applications) as well as the fact that, as per GC, there is generally no clean mapping between a scientific hypothesis (or theory) on one hand and a statistical hypothesis on the other hand with the latter often being one of many possible particular and concrete operationalizations of the former. Nonetheless, GC are correct that "there is a demand for hypothesis testing": applied researchers want to accept and reject hypotheses (and theories) and are not content with admonitions that they may only "retain the null" or that "rejection of the null should not imply acceptance of the alternative." A closer mapping between the scientific hypothesis and its operationalization as a statistical hypothesis as well as using a "default 'null model' [that] is a rich and complex model" (LS) may help in this regard where it is possible.

An additional theme concerned the notion that "probability is not a decision" (WMB)—beliefs are not actions—and the fact that "we have fallen prey to this accept-reject silliness (i.e., dichotomous thinking) because many decisions are binary" (DAB). However, as rightly pointed out by LS, when faced with a decision, the proper course of action is not to make it based on statistical hypotheses or probabilities alone but rather to conduct a full decision analysis that accounts for the costs and benefits of the various alternatives and to choose the one with, for example, the greatest expected value (although see Diaconis (2003) for a humorous cautionary note on conducting decision analyses); while, as per LS, there will still be "near hits and near misses" when using a decision analysis, a decision analysis nonetheless constitutes a major improvement over using the outcome of a statistical hypothesis test alone as the decision. In this regard, the point made by WMB that decisions are relative to person and situation and that a probability model that is useful for a given person in a given situation can be irrelevant to another person in another situation is important to bear in mind.

We further note that, while we agree with DAB that "many decisions are binary" (or at least categorical) in nature and consequently with LS that "decisions do arise that cannot be made continuously," we urge caution in this matter as many decisions that appear on the surface as binary or categorical are actually—or can be reframed to be—continuous. For example, a decision about whether or not to invest in some project can be viewed as a decision about how much to invest in the project. We believe such a continuous view of the underlying decision will naturally lead to a more continuous view of the evidence and make the issue of near hits and near misses less relevant.

Finally, while, as DAB notes, *p*-values may not cause "much harm if the focus is the primary endpoint from a protocol and the *p*-value is calculated based on a prospective analysis of that endpoint," all discussants brought up that fact that multiple comparisons—including multiple potential comparisons or the "garden of forking paths" (Gelman and Loken 2014)—are the norm in applied research and the consequence that—strictly speaking—this in practice invalidates all *p*-values except those from studies with preregistered protocols and data analysis procedures; as DAB put it, "a *p*-value has no inferential role outside the rigidness of a protocol" (and, we note, it may not inside if the underlying model that generated the *p*-value is misspecified in an important manner; we further note that while we view preregistration as often laudable, it has several limitations including being typically confirmatory and possible only in certain applied domains). This led to a discussion of alternative methods including posterior predictive probabilities of observables (WMB), hierarchical modeling and penalized (or regularized) inference techniques (GC), and false discovery rate methods (LS). We agree that all of these methods constitute a large improvement on the rote and recipe-like application of NHSTs but share the concern expressed by LS that "most applied researchers and many statisticians are not using these new tools to their full potential."

3. Social Psychology and Consumer Behavior Research

While we share the discussants' enthusiasm for recent methodological developments, we do question their applicability to the domain in which we most often work, namely, social psychology and consumer behavior. In this domain, the fundamental unit of analysis is the individual study, and the prototypical study follows a two-by-two between-subjects design where interest centers on demonstrating multiple effects—both null and nonnull—by using the linear model to conduct NHSTs on contrasts of the means of the individual-level observations in each condition. In the best of cases (even if this is not all that common), the study measures a single dependent measure and both the contrasts of interest and the data analysis procedures (e.g., outlier exclusion rules, covariates to be included in the analysis) are specified in advance.

As can be seen, dichotomization is rife in this paradigm. Not only are there the aforementioned dichotomy of the null hypothesis versus the alternative hypothesis; the dichotomization of results into the different categories statistically significant and not statistically significant; and the dichotomous thinking about there being an effect or no effect when such effects are contrasts of means, but also there is dichotomization built into the very experimental design: each experimental factor is manipulated in a dichotomous manner as if it were a light being switched on and off.

Beyond dichotomous thinking, the NHST paradigm causes additional problems in this domain. For example, because individual-level measurements are typically quite errorful, sample sizes are not especially large, and effects are small and variable, study estimates are themselves often rather noisy; noisy estimates in combination with the fact that the publication process typically screens for statistical significance results in published estimates that are biased upward (potentially to a large degree) and often of the wrong sign (Gelman and Carlin 2014). Further, the screening of estimates for statistical significance by the publication process to some degree almost encourages researchers to conduct studies with errorful measurements and small sample sizes because such studies will often yield one or more statistically significant results. Of course, all of these issues are further compounded when researchers engage in multiple comparisons—whether actual or potential.

Nonetheless, as GC noted, "there is a demand for hypothesis testing" in this domain to demonstrate effects ("to establish stylized facts" in the language of Gelman (2017)). Unfortunately, these effects are typically demonstrated by rejecting the straw man null hypothesis of zero effect decried by LS; however, it is unclear whether the rich and complex null models LS favor are possible or realistic for this data. Further, it is also unclear whether recent methodological developments can play much of a role because, for example, researchers seldom have observables for which they seek posterior probabilities, studies have no hierarchical structure, and adjustment for multiplicities via penalized inference techniques or false discovery rate methods makes little sense when zero effects are generally implausible (in this domain, there are not a small number of large effects coupled with a large number of zero effect but rather a large number of small and variable effects).

Consequently, we have been developing and encouraging the use of methods that concord with GC's call for "a greater acceptance of uncertainty and embracing of variation" while simultaneously satisfying researchers' demand to demonstrate effects. One particular area of focus has been attempting to divert attention away from individual studies, which as noted above can often be noisy, by developing meta-analytic methods (i.e., hierarchical models) that are specially tailored to the single paper meta-analysis of the multiple studies of a common phenomenon that appear in a typical research paper (McShane and Böckenholt 2017) as well as the more traditional meta-analysis of multiple studies from multiple papers that vary considerably in terms of their dependent measures and moderators (i.e., experimental factors) (McShane and Böckenholt 2017). As per GC, these methods assess and account for—indeed embrace—the variation (or heterogeneity) across multiple studies and papers (including differing degrees of variation across various dependent measures) as well as the covariation induced by the fact that observations are nested within, for example, papers, studies, groups of subjects, and study conditions; *inter alia*, this can help encourage the careful consideration of potential moderators of this variation thereby resulting in deeper and richer theories.

Further, these methods are, as noted, capable of satisfying researchers' demand to demonstrate effects, in particular via meta-analytic NHSTs. However, they do so in a perhaps subversive manner: because zero effects are generally implausible in this domain and because meta-analyses generally have much greater power than single studies, meta-analytic NHSTs are highly likely to be rejected. If the rejection of these meta-analytic NHSTs can satisfy researchers' demand to demonstrate effects, this should help divert attention away from noisy single-study NHSTs (and perhaps NHSTs in general) and free it up to focus on, for example, the estimation of effect sizes and their convergence and divergence (i.e., variation) across studies and papers as well as various dependent measures. It may also lessen considerably the degree to which the publication process screens for statistical significance (at least at the level of the individual study).

Given that the demand to demonstrate effects and the dominance of the prototypical study design are both at present firmly entrenched in this domain, we believe these methods provide researchers a means of accepting uncertainty and embracing variation that is also respectful of and responsive to their goals and data. We also believe these methods—along with other measures such as more precise individual-level measurements, larger sample sizes, a greater use of within-subjects (or longitudinal) study designs, and deeper connection between theory, measurement, and data (Gelman 2017)—should also help with current difficulties in replication.

References

Berry, D. A. (2017), "A *p*-Value to Die For," *Journal of the American Statistical Association*, 112, this issue. [904]

Briggs, W. M. (2017), "The Substitute for *p*-Values," *Journal of the American Statistical Association*, 112, this issue. [904]

Diaconis, P. (2003), "Problem of Thinking Too Much," *Bulletin of the American Academy of Arts and Sciences*, Spring 2003, 26–38. [906]

Gal, D., and Rucker, D. D. (2011), "Answering the Unasked Question: Response Substitution in Consumer Surveys," *Journal of Marketing Research*, 48, 185–195. [905]

Gelman, A. (2017), "The Failure of Null Hypothesis Significance Testing when Studying Incremental Changes, and What To Do About It," *Personality and Social Psychology Bulletin*, forthcoming. [907]

Gelman, A., and Carlin, J. (2014), "Beyond Power Calculations Assessing Type s (sign) and Type m (Magnitude) Errors," *Perspectives on Psychological Science*, 9, 641–651. [907]

——— (2017), "Some Natural Solutions to the *p*-Value Communication Problem—and Why They Won't Work," *Journal of the American Statistical Association*, 112, this issue. [904]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science," *American Scientist*, 102, 460–465. [906]

Laber, E. B., and Shedden, K. (2017), "Comment: Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-values for Statisticians," *Journal of the American Statistical Association*, 112, this issue. [904]

McShane, B. B., and Böckenholt, U. (2017), "Single Paper Meta-Analysis: Benefits for Study Summary, Theory-Testing, and Replicability," *Journal of Consumer Research*, 43, 1048–1063. [907]

——— (2017), "Multilevel Multivariate Meta-Analysis With Application to Choice Overload," *Psychometrika*. [907]

McShane, B. B., and Gal, D. (2016), "Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence," *Management Science*, 62, 1707–1718. [905]

——— (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, this issue. [904]