Routledge
Taylor & Francis Group

Basic and Applied Social Psychology

ISSN: 0197-3533 (Print) 1532-4834 (Online) Journal homepage: http://www.tandfonline.com/loi/hbas20

Four Requirements for an Acceptable Research Program

Maarten Marsman, Alexander Ly & Eric-Jan Wagenmakers

To cite this article: Maarten Marsman, Alexander Ly & Eric-Jan Wagenmakers (2016) Four Requirements for an Acceptable Research Program, Basic and Applied Social Psychology, 38:6, 308-312, DOI: 10.1080/01973533.2016.1221349

To link to this article: http://dx.doi.org/10.1080/01973533.2016.1221349

BASIC AND APPLIED SOCIAL PSYCHOLOGY
2016, VOL. 38, NO. 6, 308–312
http://dx.doi.org/10.1080/01973533.2016.1221349

Routledge
Taylor & Francis Group

COMMENT

Four Requirements for an Acceptable Research Program

Maarten Marsman, Alexander Ly, and Eric-Jan Wagenmakers

University of Amsterdam

In a recent article for this journal, Witte and Zenker (2016) proposed a research strategy that rests on the sequential evaluation of a point-alternative hypothesis. At first a large study is used to determine a "specific theoretical effect size" and then, in a series of follow-up studies, this specific effect size is contrasted against an effect size of zero. The authors deem this strategy "free of various deficits that beset dominant strategies (e.g., meta-analysis, Bayes-factor analysis)" and argue that its broad adoption constitutes "one way in which the confidence crisis may be overcome."

We agree with Witte and Zenker (2016) that it can be useful to test an alternative hypothesis that is constructed, in part or in whole, from earlier data (e.g., Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, & Ly, 2016). We also agree that it can be informative to take into account a sequence of studies as it unfolds over time (e.g., Scheibehenne, Jamil, & Wagenmakers, in press). In this comment, however, we focus mainly on areas of disagreement, which center on what we believe to be mistakes and omissions. First we address the mistakes and discuss how, in our opinion, Witte and Zenker fell prey to two fallacies: the power fallacy and the fallacy of the transposed conditional. Even for experienced scholars, these fallacies may be difficult to recognize. Second, we address the omissions and discuss four requirements for an acceptable research program.

The power fallacy

On repeated occasions, Witte and Zenker (2016) lament the lack of statistical power while boasting about the strength of statistical evidence. This confused interpretation of the data can be overcome by recognizing that power and evidence are inherently different concepts. Before we start, let's take for granted that the desired test is between $\mathcal{H}_0$: $\delta = 0$ versus a point-alternative $\mathcal{H}_1$: $\delta = 0.30$.

Now power is a predata concept, a metric constructed by averaging across all possible data sets that could be obtained in the envisioned experiment. A priori and *on average*—with respect to all possible data sets—experiments designed with low power are unlikely to yield a significant outcome given that $\mathcal{H}_1$ is true. In contrast, evidence is a postdata concept. In this specific scenario the evidence is given by the likelihood ratio, that is, the relative probability of the observed data under the competing hypotheses. The likelihood ratio considers only the data that have in fact been obtained.

As discussed elsewhere in detail, after the data have been observed, data that could have been observed but were not are evidentially irrelevant (e.g., Berger & Wolpert, 1988; Bayarri, Benjamin, Berger, & Sellke, 2016; Wagenmakers, Verhagen, et al., 2015; Wagenmakers et al., in press). Our predata state of knowledge has been altered by the observation of the data, and after the data have arrived, our postdata state of knowledge is all that ought to matter.

When the predata concept of power is erroneously used for postdata purposes—such as inference and the quantification of evidence—this entails a deliberate loss of important information, namely, the actual outcomes of the experiment.

The fallacy of the transposed conditional

Witte and Zenker (2016) correctly point out that the Bayes factor is the probability of the data under $\mathcal{H}_0$ versus $\mathcal{H}_1$ (Wagenmakers, Morey, & Lee, 2016). They also acknowledge that the Bayes factor and the likelihood ratio are "quantitatively" equivalent whenever the hypotheses are both simple (i.e., consisting of a single specified point value for effect size). However, Witte and Zenker argue that changing the nomenclature—from Bayes factors to likelihood ratios—allows one to interpret the likelihood ratio as the relative plausibility of the hypotheses. So even though what is calculated is the relative probability of the data given the hypotheses, the result is interpreted as the relative probability of the

CONTACT Maarten Marsman m.marsman@uva.nl Psychological Methods Department, University of Amsterdam, P.O. Box 15906, 1001 NK, Amsterdam, the Netherlands.

hypotheses given the data. By doing so, Witte and Zenker commit the fallacy of the transposed conditional.

Unfortunately, in statistical inference there is no such thing as a free lunch (Rouder, Morey, Verhagen, Province, & Wagenmakers, in press). Any time one wishes to assign probabilities to parameters or models, one is automatically committed to the Bayesian framework (Ly, Verhagen, & Wagenmakers, 2016a, 2016b). Specifically, the only way to obtain a posterior probability is by using the data to update a prior probability. Bayes factors quantify the extent to which the data change the prior model odds to posterior model odds, and as such they can be considered the relative evidence that the data provide for the models under consideration. The Bayes factor is therefore only one ingredient for inference. The other ingredient is the prior model odds. One is licensed to interpreted Bayes factors (or likelihood ratios, for simple models) as posterior odds, but only when the prior odds equals 1, and *not* when the prior odds is ignored because it makes researchers uncomfortable, like a family member who has unexpectedly decided to vote for Donald Trump.

To appreciate the importance of the prior odds, consider the competing models $\mathcal{H}_1$: "people have extrasensory perception (ESP)" versus $\mathcal{H}_0$: "people do not have ESP." Few researchers would seriously entertain equal prior odds in this case. Moreover, suppose the likelihood ratio for an ESP experiment yielded a factor of 30 in favor of ESP; do we conclude from this that the ESP hypothesis is 30 times more likely than the null hypothesis? Of course we do not, and if the authors' methodology were to sanction this inference (which it does not), then this would be a compelling argument against their methodology instead of a compelling argument for ESP. Extraordinary claims require extraordinary evidence, and in order to assess the posterior plausibility of ESP one needs to combine the evidence from the data (i.e., the Bayes factor) with the prior plausibility of the ESP phenomenon (Wagenmakers, Wetzels, et al., 2015).

Requirements of a research program

A research program that can alleviate the current "crisis of confidence" (Pashler & Wagenmakers, 2012) needs to be more ambitious than the approach proposed by Witte and Zenker (2016). Next we outline four key requirements and point to the relevant literature.

I. Preregistration

Philosophers, psychologists, physicians, and physicists have long argued that empirical research needs to respect the distinction between work that is exploratory or hypothesis generating and work that is confirmatory or hypothesis testing, and that this needs to be done by preregistering the analysis plan in all of its details (e.g., Barber, 1976; Chambers, 2013; Feynman, 1998; Goldacre, 2009; Peirce, 1878, 1883; Wagenmakers, Wetzels, Boorsboom, van der Maas, & Kievit, 2012).

These theoretical arguments have garnered empirical support in the sense that preregistered replications rarely support the original effects (e.g., Nosek & Lakens, 2014; Open Science Collaboration, 2012). Without preregistration, researchers can easily and unwittingly fall prey to hindsight bias and confirmation bias. In our opinion, any research program that does not include preregistration is seriously incomplete.

II. Transparency

In reproducible research, transparency is essential. Indeed, one can argue that preregistration falls under the general heading of transparency as well. Here we use transparency to refer to open materials, open data, and open analysis code. Recent initiatives such as TOP (Transparency and Openness Promotion; Nosek et al., 2015), PRO (The Peer Reviewers' Openness Initiative; Morey et al., 2016), and the Center for Open Science badges for good academic behavior (Kidwell et al., 2016) aim to change the dominant culture so that openness becomes the norm instead of the exception.

In our own work, we have developed the open-source statistical software program JASP (jasp-stats.org; JASP Team, 2016). In JASP, users can save data, analysis input, analysis output, and analysis annotations in a single .jasp file.[1] When this file is uploaded to the Open Science Framework, the OSF JASP previewer allows anybody with an online browser to inspect the annotated output, even without having JASP installed. An example featuring Adam Sandler is available at https://osf.io/3pbzk/ (Wagenmakers, Morey, & Lee, 2016).

III. Comprehensive knowledge updating

A mature research program allows knowledge to be updated as new data come in (Scheibehenne et al., in press). This requirement is fulfilled by Witte and Zenker (2016), but only in part; what is updated is the likelihood ratio, but not the value of the parameter. In other words, based on the initial study, Witte and Zenker committed themselves 100% to the single point estimate $\delta = 0.30$. This violates what Lindley termed "Cromwell's rule." Cromwell famously told the Church of Scotland, "I beseech you, in the bowels of Christ, think it possible you may be mistaken." Cromwell's rule states that one should not categorically rule out anything, for this makes

it impossible to learn. As explained by Lindley (1985), "So leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved" (p. 104).

Occasionally there are reasons to violate Cromwell's rule. For instance, one may wish to evaluate the relative adequacy of the predictions from a theoretically meaningful hypothesis—perhaps a general law or invariance (Rouder Speckman, Sun, Morey, & Iverson, 2009), or perhaps a physical law involving gravity or the speed of light. In the current example, however, the point estimate of 0.30 is devoid of theoretical content; the effect size could differ from one context to the next, or it could be lower or higher. The original data set suggested $\delta = 0.30$, but what if a second, much larger data set,[2] had suggested $\delta = 0.10$? This value is still consistent with the general theory of there being an effect, only it is a little smaller than suggested in the original study. The likelihood ratio would have favored $\mathcal{H}_0$, but at the same time it would be obvious that $\mathcal{H}_0$ is false. This is the equivalent of the Lindley's astronaut scenario.

The correct way to update knowledge is to update both the plausibility of competing models and the plausibility of the parameters within those models.[3] This implies that we also need priors on the parameters within the models. Those who feel uneasy about injecting "subjective" knowledge into their statistical analyses may be comforted by the results from Bickel and Kleijn (2012), Kleijn and van der Vaart (2012), and van der Vaart (1998), who showed that the influence of the prior on the posterior washes out quickly, especially for the regular models used in the psychological sciences.[4]

Bayesian updating entails that the posterior distribution after study n becomes the prior distribution for the analysis of study $n + 1$ (Ly, Etz, & Wagenmakers, 2015; Verhagen & Wagenmakers, 2014; Wagenmakers, Verhagen, & Ly, 2016). Under the assumption that the different studies are measuring the same effect, this method of knowledge updating is internally consistent and implements a principled method of scientific learning.[5]

IV. Acknowledging uncertainty

In our experience, researchers strongly desire unambiguous yes/no answers, even when these are unavailable due to the stochastic nature of the data. Paradoxically, the noisier the data, the stronger this desire seems to become.

The decision-making framework of null-hypothesis significance testing offers some certainty: If $p < .05$, we may "reject the null hypothesis." This is fulfilling, because by making a decision we have swept all of the existing uncertainty under the rug. There is no more need to debate the outcome any longer, the researcher may feel, because we were sanctioned to make a Decision and Reject the Null Hypothesis. After the Gordian knot has been cut, it is futile to argue about other possible decisions that could have been made. This way, null-hypothesis significance testing offers an illusion of certainty, and with it the protection again critique and self-doubt.

Unfortunately there are several problems with the decision-making framework of null-hypothesis significance testing. The list is endless, but here we highlight the following concerns:

1. Utilities are ignored. If the purpose of statistical inference in academia is to make decisions, then one needs to specify utilities or loss functions associates with the potential outcomes (e.g., Lindley, 1985). Without utilities there can be no sensible decision making.
2. Scientists often do not make decisions. As stated by Rozeboom (1960), "The null-hypothesis significance test treats acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data—i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested*" (p. 420).
3. The p value from the framework of null-hypothesis significance testing—upon which the Decision to Reject the Null Hypothesis is based—is "violently biased against the null hypothesis" (Edwards, 1965, p. 400); see also Berger & Delampady, 1987; Edwards, Lindman, & Savage, 1963; Johnson, 2013; Marsman & Wagenmakers, in press; Sellke, Bayarri, & Berger, 2001; Wetzels et al., 2011). For these and other reasons we sympathize with the p value ban in *Basic and Applied Social Psychology* (Trafimow & Marks, 2015).

Instead of using ad hoc decision rules for seeking certainty where there is none, it is better to acknowledge and quantify uncertainty. If a Bayes factor indicates that the data are 4 times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$, this does not mean that $\mathcal{H}_0$ has been refuted, or that $\mathcal{H}_1$ is true. Authors should make claims that are in accordance with the strength of evidence in the data—often, this means that the claims should be more modest. In turn, editors and reviewers should reward such modesty, not punish it.

Concluding comments

We proposed four requirements for an acceptable research program, which we believe to be at odds with Witte and Zenker's proposal. Specifically, their proposal fails to acknowledge uncertainty and does not result in coherent knowledge updates. Witte and Zenker sweep prior model probabilities under the rug and construct a point alternative hypothesis from noisy data.

For comprehensive knowledge updating, that is, for *statistical learning*, one has to adhere to the laws of probability, the same way the motion of the stars follows the laws of physics. Our advocacy for Bayesian methods in psychology is, in essence, a call to adopt a principled method of learning. This call is neither new (Edwards et al., 1963) nor controversial, as Bayesian methods have been adopted with great success in many fields. In psychology, Bayesian methods have traditionally been used to assess people's ability for "optimal information processing." However, the same researchers who claim that optimal information processing requires Bayes's rule will resist applying that rule when they themselves have to process information; instead, they happily report p values.[6]

The reward for adopting Bayesian methods in psychology is substantial: Not only do our conclusions respect the laws of probability, but our uncertainty is automatically quantified in terms of posterior distributions. These posteriors quantify our complete knowledge and can be transformed into so-called posterior predictives that give an indication of how our previous findings might generalize to new experiments (Liu & Aitkin, 2008). As a measure of replicability, the posterior predictive will outperform Witte and Zenker's (2016) approach. This is due to their over enthusiastic commitment to a single point alternative hypothesis that disregards uncertainty, making predictions overly confident (Aitchison & Dunsmore, 1975).

The proposed four requirements for an acceptable research program are relatively straightforward to execute, but they imply that researchers acknowledge and counteract fundamental human biases and desires. Implementing the program therefore requires a change in academic culture. Academic culture is difficult to change, but the past 5 years have demonstrated that it can be done. Driven by the combined efforts from researchers, journals, funders, and institutes (especially the Center for Open Science), there has been a dramatic and positive reorientation of academic values. The caterpillar known as psychological science has finally started its metamorphosis, and our bet is on a Bayesian butterfly.

Notes

1. The analysis output may also be saved separately.
2. For concreteness and to avoid ambiguity, let's say 1,000 times as large.
3. Point-hypotheses are a good approximation to posterior distributions that are highly peaked, but in the case of Witte and Zenker (2016) we see no compelling reason in this case to violate Cromwell's rule and update knowledge only partially.
4. Note that the effect of the likelihood often dwarfs the effect of the prior.
5. When the studies are measuring a different but similar effect, then one may apply a hierarchical model (Shiffrin, Lee, Kim, & Wagenmakers, 2008).
6. What explains the apparent reluctance of psychologists to use the rule they know is optimal? We believe it might be utility—classical analyses are easy to report and may be thought to raise fewer questions from the reviewers. Such considerations suggest that it may be Bayesian to refrain from reporting a Bayesian analysis.

Funding

This work was supported by the starting grant "Bayes or Bust" awarded by the European Research Council (Grant number 283876).

References

Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge, UK: Cambridge University Press.

Barber, T. X. (1976). *Pitfalls in human research: Ten pivotal points*. New York, NY: Pergamon Press Inc.

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle* (2nd ed.). Hayward, CA: Institute of Mathematical Statistics.

Bickel, P. J., & Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, *40*, 206–237.

Chambers, C. D. (2013). Register reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

Edwards, W. (1965). Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, *63*, 400–402.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Feynman, R. (1998). *The meaning of it all: Thoughts of a citizen–scientist*. Reading, MA: Perseus Books.

Goldacre, B. (2009). *Bad science*. London, UK: Fourth Estate.

JASP Team, T. (2016). JASP (Version 0.7.5.6)[Computer software].

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 19313–19317.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low cost, effective method for increasing transparency. *PLoS Biology*, *14*, e1002456.

Kleijn, B., & van der Vaart, A. (2012). The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, *6*, 354–381.

Lindley, D. V. (1985). *Making Decisions* (2nd ed.). London, UK: Wiley.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362–375.

Ly, A., Etz, A., & Wagenmakers, E. J. (2015). *Replication Bayes factors*. Manuscript in preparation.

Ly, A., Verhagen, A. J., & Wagenmakers, E. J. (2016a). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Ly, A., Verhagen, A. J., & Wagenmakers, E. J. (2016b). Harold Jeffreys's Default Bayes Factor Hypothesis Tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Marsman, M., & Wagenmakers, E. J. (in press). Three insights from a Bayesian Interpretation of the one–sided *p* value. *Educational and Psychological Measurement*.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., & Zwaan, R. A. (2016). The peer reviewers' openness initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*, 150547.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., & Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348, 1422–1425.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141.

Open Science Collaboration, T. (2012). An open, large–scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660.

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

Peirce, C. S. (1878). Deduction, induction, and hypothesis. *Popular Science Monthly*, *13*, 470–482.

Peirce, C. S. (1883). A theory of probable inference. In C. S. Peirce (Ed.), *Studies in logic*, pp. 126–181. Boston, MA: Little & Brown.

Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E. J. (in press). Is there a free lunch in inference? *Topics in Cognitive Science*.

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*, 225–237.

Rozeboom, W. W. (1960). The fallacy of the null–hypothesis significance test. *Psychological Bulletin*, *57*, 416–428.

Scheibehenne, B., Jamil, T., Wagenmakers, E. J. (in press). Bayesian evidence synthesis can reconcile seemingly inconsistent results: The case of hotel towel reuse. *Psychological Science*.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician*, *55*, 62–71.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*, 1–2.

van der Vaart, A. W. (1998). *Asymptotic statistics*. New York, NY: Cambridge University Press.

Verhagen, A. J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E. J., Verhagen, A. J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426.

Wagenmakers, E. J., Verhagen, A. J., Ly, A., Bakker, M., Lee, M. D., Matzke, D. & Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, *47*, 913–917.

Wagenmakers, E. J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York, NY: Wiley and Sons.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Kievit, R., & van der Maas, H. L. J. (2015). A skeptical eye on psi. In E. May & S. Marwaha (Eds.), *Extrasensory perception: Support, skepticism, and science*, (pp. 153–176). ABC-CLIO.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science*, *6*, 291–298.

Witte, E. H., & Zenker, F. (2016). Reconstructing recent work on macro-social stress as a research program. *Basic and Applied Social Psychology*, *38*(6), 301–307. doi:10.1080/01973533.2016.1207077