General Article

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(3) 367–374

DOI: 10.1177/2515245918779348
www.psychologicalscience.org/AMPPS

# Bayesian Reanalyses From Summary Statistics: A Guide for Academic Consumers

**Alexander Ly[1,2], Akash Raj[1], Alexander Etz[3], Maarten Marsman[1], Quentin F. Gronau[1], and Eric-Jan Wagenmakers[1]**
[1]Psychological Methods Department, University of Amsterdam; [2]Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam; and [3]Department of Cognitive Sciences, University of California, Irvine

**Abstract**

Across the social sciences, researchers have overwhelmingly used the classical statistical paradigm to draw conclusions from data, often focusing heavily on a single number: *p*. Recent years, however, have witnessed a surge of interest in an alternative statistical paradigm: Bayesian inference, in which probabilities are attached to parameters and models. We feel it is informative to provide statistical conclusions that go beyond a single number, and—regardless of one's statistical preference—it can be prudent to report the results from both the classical and the Bayesian paradigms. In order to promote a more inclusive and insightful approach to statistical inference, we show how the Summary Stats module in the open-source software program JASP (https://jasp-stats.org) can provide comprehensive Bayesian reanalyses from just a few commonly reported summary statistics, such as *t* and *N*. These Bayesian reanalyses allow researchers—and also editors, reviewers, readers, and reporters—to (a) quantify evidence on a continuous scale using Bayes factors, (b) assess the robustness of that evidence to changes in the prior distribution, and (c) gauge which posterior parameter ranges are more credible than others by examining the posterior distribution of the effect size. The procedure is illustrated using Festinger and Carlsmith's (1959) seminal study on cognitive dissonance.

**Keywords**

Bayes factor, Bayesian inference, data visualization, effect size, effects, hypothesis testing, *p* value, cognitive dissonance, open materials

Received 6/2/17; Revision accepted 3/27/18

Classical null-hypothesis significance testing (NHST) allows researchers to evaluate scientific propositions in a seemingly straightforward manner: Whenever the *p* value falls below a threshold $\alpha$ (usually set to .05), researchers feel licensed to reject the null hypothesis ($H_0$) that the effect is absent and to embrace the alternative hypothesis ($H_1$) that the effect is present. For example, in the results section, one may encounter conclusions such as "overall classification accuracy was greater than chance," "the analysis revealed a main effect of the manipulation," and "the correlation was significant"; in the discussion section, these statements typically are abstracted even further from the standard NHST framework, conveying the impression that whenever *p* is below the .05 threshold, the data strongly favor $H_1$ over $H_0$ (i.e., no effect).

The social sciences' mechanistic use of *p* values appears to be at odds with the recent warning issued by the American Statistical Association (ASA; Wasserstein & Lazar, 2016): "The widespread use of 'statistical significance' (generally interpreted as '$p \leq .05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process" (p. 131). Indeed, *p* values have been critiqued on numerous grounds (e.g., Greenland et al., 2016; Nickerson, 2000; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Wagenmakers, Marsman, et al.,

**Corresponding Author:**
Alexander Ly, University of Amsterdam, Psychological Methods Department, Postbus 15906, 1001 NK Amsterdam, The Netherlands
E-mail: a.ly@uva.nl

2018). One widely appreciated concern is that *p* values do not convey information about the size of the effect or the precision with which that effect is estimated (e.g., Cumming, 2014).

There is a growing trend for psychologists to employ Bayesian statistics (e.g., Vandekerckhove, Rouder, & Kruschke, 2018; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017; Wagenmakers, Morey, & Lee, 2016), a prominent alternative to *p*-value NHST. Within the Bayesian framework, prior uncertainties about parameters and models are updated by means of observed data to yield posterior uncertainties (Etz, Gronau, Dablander, Edelsbrunner, & Baribault, 2018). For instance, the posterior distribution quantifies one's knowledge about a nonzero effect size, which is useful for parameter estimation; on the other hand, the Bayes factor contrasts the predictive adequacy of two competing models, which is useful for hypothesis testing (for technical details, see Etz & Vandekerckhove, 2018). Specifically, the Bayes factor quantifies the degree to which the data are more likely under one model versus another (e.g., Etz & Wagenmakers, 2017; Jeffreys, 1961; Kass & Raftery, 1995; Myung & Pitt, 1997). For example, the Bayes factor can be used to compare the hypothesis that there is no effect ($H_0$) with the hypothesis that there is an effect ($H_1$). The value of the Bayes factor in favor of $H_1$ over $H_0$, denoted as $\text{BF}_{10}$, has an intuitive interpretation; a $\text{BF}_{10}$ of 7 indicates that the observed data are 7 times more likely under $H_1$ than under $H_0$, whereas a $\text{BF}_{10}$ of 0.2 indicates that the observed data are 5 times more likely under $H_0$ than under $H_1$. In general, $\text{BF}_{10}$ is a nonnegative number, and higher values indicate that the data provide more evidence for $H_1$ over $H_0$, whereas lower values indicate that the data provide less evidence for $H_1$ over $H_0$. When $H_1$ and $H_0$ are equally plausible a priori, a $\text{BF}_{10}$ of 6 indicates that the posterior plausibility for $H_1$ is 86% (i.e., 6/7), leaving 14% (i.e., 1/7) posterior probability for $H_0$.

A discussion on the merits and demerits of the different statistical paradigms is beyond the scope of this article. We agree with the ASA's recommendation to go beyond *p*, and we believe that it is prudent to adopt an inclusive statistical approach. When the results of different statistical paradigms point in the same direction, this bolsters one's confidence in the conclusions, but when the results are in blatant contradiction, this weakens one's confidence.

In the spirit of promoting a more inclusive statistical approach, our primary goal in this article is to demonstrate the ease with which published classical results can be subjected to a Bayesian reanalysis using the recently developed Summary Stats module in JASP (JASP Team, 2017). Depending on the analysis at hand, this module takes as input a commonly reported statistic such as *t, r*, or *R* together with sample size *N* and returns a comprehensive Bayesian assessment.[1] An important point is that this Bayesian assessment can be executed in the absence of the raw data, which is essential when the data are no longer available or when they cannot be shared; but even when the raw data are publicly available, the reanalysis with the Summary Stats module is much more efficient—reviewers, readers, and reporters can obtain a comprehensive Bayesian assessment almost instantaneously. We believe that the richness of a Bayesian report contrasts favorably with a report of just the summary statistics themselves. We illustrate this claim using a seminal study published more than half a century ago.

## Disclosures

The analyses presented here are available via the Open Science Framework (file named festingerCarlsmith1959.jasp), at https://osf.io/7t2jd/.

## Festinger and Carlsmith's (1959) Cognitive Dissonance Study

In a landmark publication (cited more than 3,540 times as of February 2, 2018, according to Google Scholar), Festinger and Carlsmith (1959) outlined a theory to account for *cognitive dissonance*, a phenomenon they described as follows:

> If a person is induced to do or say something which is contrary to his private opinion, there will be a tendency for him to change his opinion so as to bring it into correspondence with what he has done or said. (p. 209)

In earlier experiments on cognitive dissonance (e.g., Kelman, 1953), participants were induced to make a statement contrary to their personal opinion for a chance to gain a reward. It was hypothesized that the opinion would change more when the potential reward was greater, but the data showed the reverse: The smaller the reward, the greater the change in opinion. Festinger and Carlsmith proposed a theory that could account for this behavioral pattern, and they subsequently put that theory to the test in an ingenious experiment.

Festinger and Carlsmith's (1959) experiment included control, high-reward, and low-reward conditions, each with 20 participants. All participants performed a boring task for 1 hr, after which they were asked to take a survey and answer questions about, among other things, their enjoyment of the study. Where the conditions differed was in what happened after participants completed the boring task but before they completed the survey. In the reward conditions, participants were asked to tell a confederate that the experiment was interesting and fun; in return, they received either \$20 (high reward) or \$1 (low reward). In the control condition, participants went straight to the survey and did not interact with the confederate. According to Festinger and Carlsmith, the crucial test of their theory lay in comparing the enjoyment ratings from the low-reward condition and the high-reward condition; the low-reward condition was predicted to yield higher enjoyment ratings. Results were in line with the theory's prediction: The (sample) mean enjoyment rating was higher in the low-reward group than in the high-reward group, $t(38) = 2.22$, $p = .032$, and Festinger and Carlsmith took this as support for their theoretical position. No effect-size estimate was reported in the original publication, but the effect size can be easily computed from the *t* value and the group sizes, which yield a Cohen's *d* of 0.702.

## Bayesian Reanalysis

If Festinger and Carlsmith's (1959) study were published today and we wished to conduct a Bayesian reanalysis of their result, then we would ask them to share their data with us. Unfortunately, this study was published six decades ago, and the raw data are no longer available. However, the Summary Stats module in JASP affords a comprehensive Bayesian reanalysis of the experiment using only the test statistic reported in the original publication.[2] Inputting the reported *t* value and sample sizes for the two groups yields the results shown in Figure 1.

In Figure 1a, the dashed line represents the default prior distribution for the population effect size, $\delta$, under $H_1$: a zero-centered *Cauchy* distribution (i.e., a *t* distribution with 1 degree of freedom; Jeffreys, 1948; Ly, Verhagen, & Wagenmakers, 2016a, 2016b), with a default scale, $\gamma$, of 0.707 (e.g., Morey & Rouder, 2015). Thus, under $H_1$—that is, presuming the effect is present—the expectation is that the effect is most likely to be small, although the possibility that it is large is not ruled out.

The solid line in Figure 1a is the posterior distribution for effect size, that is, the knowledge about effect size obtained after updating the prior distribution using the observed data, and presuming that $H_1$ holds. This posterior distribution of $\delta$ has a median of 0.571[3] and a relatively wide 95% credible interval that ranges from –0.032 to 1.197. In other words, 95% of the posterior mass lies in the interval from –0.032 to 1.197; clearly, the effect has not been estimated with much precision. More generally, by computing the area under the posterior distribution between $\delta = a$ and $\delta = b$, one can assess how plausible it is that the population effect size falls in the interval from *a* to *b* after the data have been observed (e.g., Wagenmakers, Love, et al., 2018; Wagenmakers et al., 2016). For instance, by comparing the area under the posterior distribution to the right of zero against the area under the posterior distribution to the left of zero, one quantifies how much more likely it is that the effect is positive rather than negative, under $H_1$—that is, under the presumption that the effect is present.

In general, the posterior distribution quantifies all that one knows about the population effect size given that $H_1$ holds and the effect exists. The latter point is worth emphasizing because it has been argued that one may perform a Bayesian null-hypothesis test by judging whether the 95% credible interval overlaps with zero. Despite its beguiling simplicity, such a procedure is incorrect (Berger, 2006; Jeffreys, 1961; Wagenmakers et al., 2017) because it begs the question: The extent to which $H_0$ is plausible cannot be assessed when this hypothesis has been ruled out in advance; that is, under the continuous prior distribution (e.g., the Cauchy prior) specified for $H_1$, the probability of any single point, such as $p(\delta = 0)$, equals zero.

In order to perform a Bayesian hypothesis test, one needs to compare the predictive performance of $H_0$ against that of $H_1$. The result of this comparison is the Bayes factor ($\mathrm{BF}_{10}$), and Figure 1a reveals that it equals 2.056 for Festinger and Carlsmith's (1959) data; that is, the observed data are only about twice as likely under $H_1$ than under $H_0$. Bayesian statistician Harold Jeffreys (1961) deemed this level of evidence "not worth more than a bare mention" (p. 432). The proportion wheel in this panel of the figure provides a visualization of the strength of the evidence (see also Wagenmakers & Gronau, 2015).

The Bayes factor quantifies relative predictive performance, and the predictive performance from $H_1$ is determined in part by the prior distribution. Under a default prior specification, it is natural to wonder how robust the conclusions are to plausible changes in the prior distribution. To address this issue, the Summary Stats module allows one to select the option "Bayes factor robustness check." Figure 1b shows the result: The value of $\mathrm{BF}_{10}$ is graphed as a function of the scale, $\gamma$, of the Cauchy prior distribution. The values on the *x*-axis range from $\gamma = 0$ (when $H_1$ reduces to $H_0$ and $\mathrm{BF}_{10}$ is 1 regardless of the data) to $\gamma = 1.5$. Across this entire range, the Bayes factor in favor of $H_1$ over $H_0$, $\mathrm{BF}_{10}$, never exceeds 3; in fact, the maximum $\mathrm{BF}_{10}$ equals 2.159, when $\gamma$ is set to 0.459.

So far, we assumed that the prior for effect size is always centered at zero. However, the Bayesian

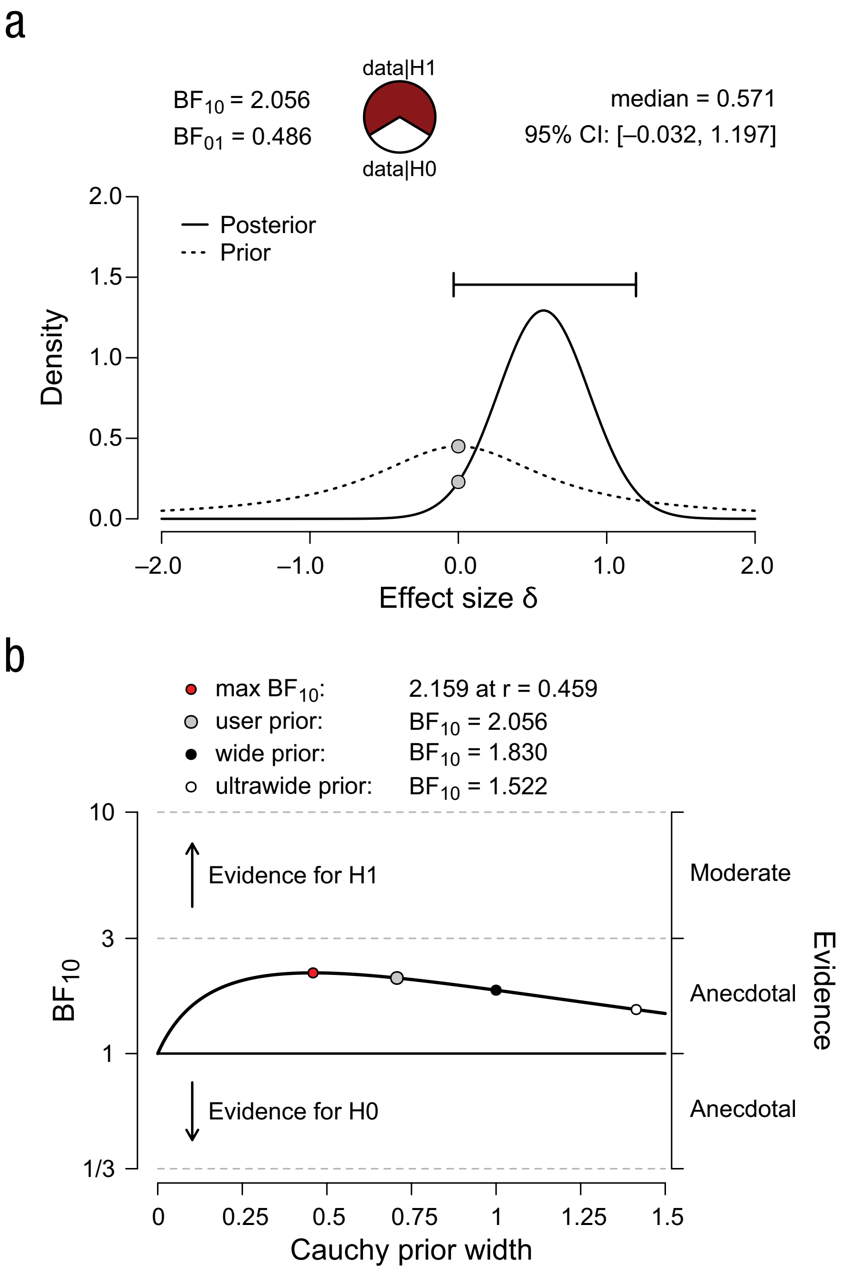

**Fig. 1.** Screenshot showing JASP's output for a comprehensive Bayesian reanalysis of the seminal study by Festinger and Carlsmith (1959), obtained by entering $t$ = 2.22 and $N_1 = N_2$ = 20 into the JASP Summary Stats module. The graph in (a) shows the default prior distribution and the posterior distribution for the population effect size, δ, under the alternative hypothesis ($H_1$). The 95% credible interval (CI) is indicated by the bar. $\mathrm{BF}_{10}$ is the Bayes factor in favor of $H_1$ over the null hypothesis ($H_0$), and $\mathrm{BF}_{01}$, which is equal to $1/\mathrm{BF}_{10}$, is the Bayes factor in favor of $H_0$ over $H_1$. The graph in (b) shows the value of $\mathrm{BF}_{10}$ as a function of the width (i.e., scale, or γ) of the Cauchy prior. For the wide prior, γ = 1, and for the ultrawide prior, γ = √2; the user prior corresponds to the scale chosen by the user, which in this case is 0.707. The arrows indicate ranges in which the $\mathrm{BF}_{10}$ indicates there is evidence for $H_1$ (i.e., values > 1) and in which the $\mathrm{BF}_{10}$ indicates there is evidence for $H_0$ (i.e., values < 1). See the text for further details.

framework can be extended to include informed prior distributions that incorporate context-specific expectations and need not be centered around zero (Gronau, Ly, & Wagenmakers, 2017). As in the reanalysis with default priors, the reanalysis with informed priors is a function solely of the summary statistics. To illustrate the use of an informed prior, we once again reanalyze Festinger and Carlsmith's (1959) study, but this time

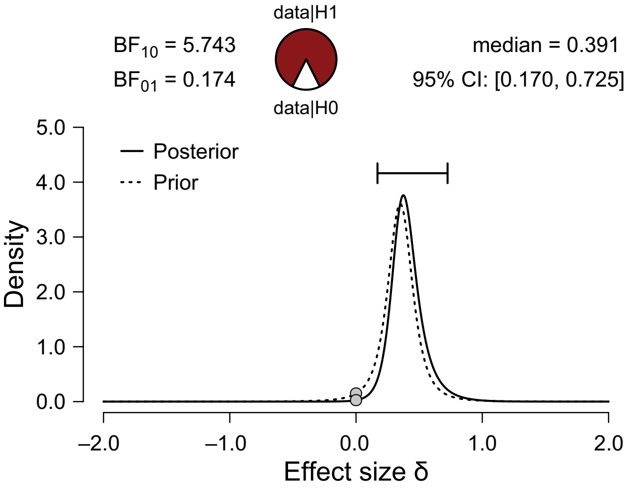

**Fig. 2.** Screenshot of JASP's output for a Bayesian reanalysis of Festinger and Carlsmith's (1959) result, obtained by entering $t = 2.22$ and $N_1 = N_2 = 20$ into the JASP Summary Stats module and changing the prior setting to a $t$ distribution with a location of 0.350, a scale of 0.102, and 3 degrees of freedom. $H_1$ = the alternative hypothesis; $H_0$ = the null hypothesis; $BF_{10}$ = Bayes factor in favor of $H_1$ over $H_0$; $BF_{01} = 1/BF_{10}$ and is the Bayes factor in favor of $H_0$ over $H_1$; CI = credible interval.

with two different priors that are centered away from zero. The first informed prior was elicited from Suzanne Oosterwijk, a social psychologist at the University of Amsterdam. This Oosterwijk prior (Gronau et al., 2017) was elicited in the context of a specific effect, but we believe it is plausible more generally for effects in experimental psychology whose presence needs to be ascertained by a statistical analysis. The *Oosterwijk prior* is a $t$ distribution with a location of 0.350, a scale of 0.102, and 3 degrees of freedom (dashed line in Fig. 2); it assigns most mass to effect sizes from 0.1 to about 0.6. Because this prior is highly informative and the number of observations in Festinger and Carlsmith's study is fairly small, the change from prior to posterior (solid line in Fig. 2) is modest in this case. The $BF_{10}$ indicates that the data are 5.743 times more likely under the informed alternative than under $H_0$. Under equal prior-model probabilities, this reanalysis leaves $H_0$ a nonnegligible posterior probability of 14.8% (i.e., 1/6.743).

One might suspect that it is possible to find as much evidence for $H_1$ as desired for a given data set just by changing the prior distribution. This, however, is not true. The second informed prior that we consider is an *oracle prior* that assigns all of its mass to a single point: the observed effect size, $d$, of 0.702. This "prior" can be obtained in JASP by choosing a normal distribution prior with mean of 0.702 and standard deviation of 0. Note that as a prior it is unrealistic, because in practice it is impossible to know the observed effect size before conducting the experiment. However, this prior showcases the maximum evidence possible in favor of $H_1$ (Edwards, Lindman, & Savage, 1963). Using this oracle prior, we obtain a Bayes factor of 10.45 in favor of $H_1$ over $H_0$. Hence, even if researchers are blatantly cheating by assigning all prior mass to the observed effect size, the data are only 10.45 times more likely under $H_1$ than under $H_0$; under equal prior-model probabilities, such an extreme form of cheating still leaves $H_0$ a posterior probability of 8.7% (i.e., 1/11.45). In this particular scenario, a seminal result, significant with a $p$ value of .032, does not yield compelling evidence against $H_0$ when assessed from a default or an informed Bayesian perspective.[4] We wish to emphasize that the strength of evidence provided by a Bayes factor can be best appreciated by considering the raw numbers, perhaps visualized as a proportion wheel (see Figs. 1 and 2); the classification scheme proposed by Jeffreys (1961) provides a useful but rough guideline that should not take precedence over a more careful assessment of the strength of evidence.

In sum, the Bayesian reanalyses shown in Figures 1 and 2 are easily obtained in JASP and paint an inferential picture more complete than the one provided by the statement "$t(38) = 2.22$, $p = .032$."

## Concluding Comments

The Summary Stats module in JASP unlocks a comprehensive Bayesian experience from a few commonly reported summary statistics. We have illustrated the use of the module for the case of an independent-samples $t$ test, but the Summary Stats module can also be used for inferences concerning paired-samples $t$ tests, correlation coefficients, binomial proportions, and linear regression models. An entire literature filled with classical statistics is now open for straightforward Bayesian reanalysis.

In addition to looking back on the existing literature, one can look forward. For instance, editors and reviewers may request that authors include a Bayesian analysis alongside the results obtained using classical statistical methods, and such a Bayesian analysis can be obtained with JASP in mere seconds. For a specific data set, the results obtained with classical and Bayesian analyses may disagree (e.g., Wetzels et al., 2011). We believe that such a discrepancy is cause for additional reflection, because it suggests that the data are perhaps not as informative as one would have otherwise believed. Reviewers and editors can request that authors acknowledge this uncertainty and be transparent about the conflicting accounts of the data.

An additional advantage of a Bayesian analysis is that one can use the data efficiently to inform follow-up

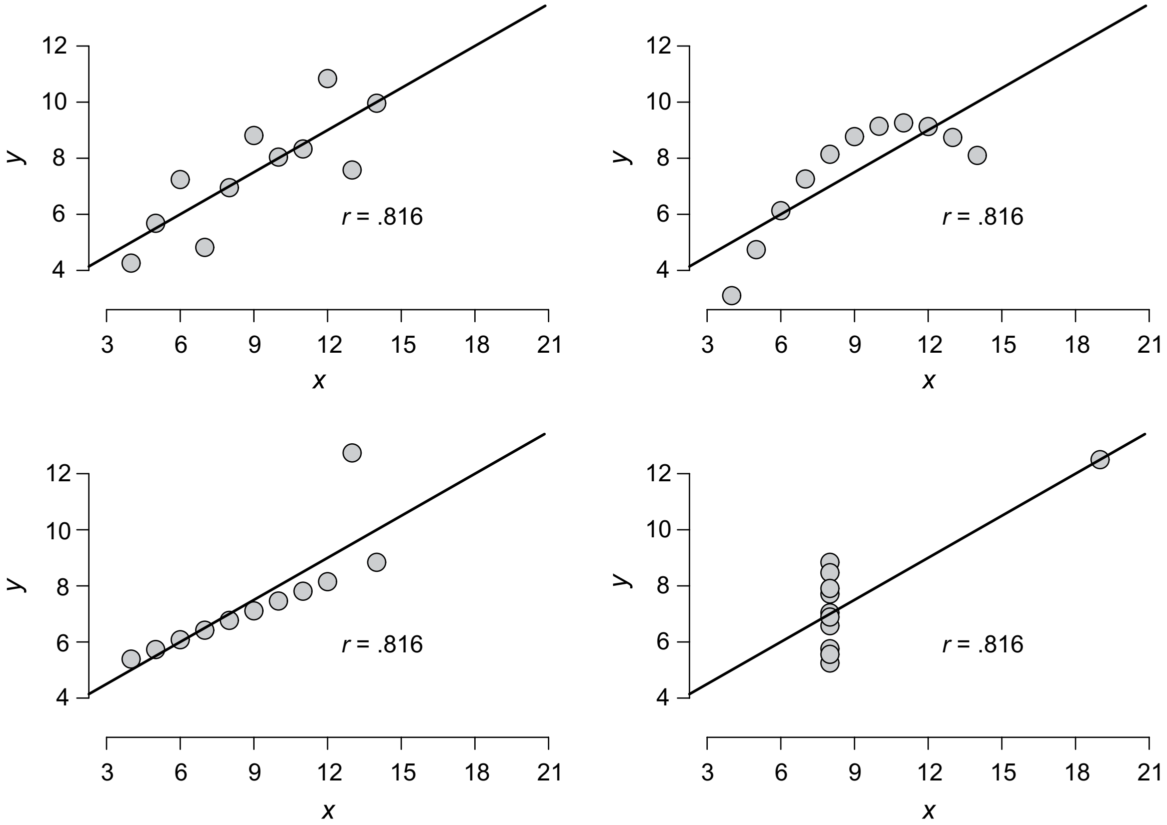

**Fig. 3.** Anscombe's (1973) quartet. The summary statistics are identical in the four panels: The *x* variable has a sample mean of 9 and a sample variance of 11, the *y* variable has a sample mean of 7.50 and a sample variance of 4.125, and the observed Pearson correlation coefficient is .816. These scatterplots illustrate the need to visualize the data, as the summary statistics alone do not tell the full story; in this case, inspection of the raw data shows that inferring a linear association is meaningful in only one of the four cases.

studies by taking the posterior distribution from the current study as a prior distribution for further studies. This allows one to compute *replication Bayes factors* (Ly, Etz, Marsman, & Wagenmakers, in press; Verhagen & Wagenmakers, 2014), which quantify the additional evidence brought forth by the new data.

Note, however, that even when the summary statistics are *sufficient* (i.e., they capture all relevant information; e.g., Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017) on general grounds, it is still beneficial to have access to the raw data. The raw data can be used to confirm that the statistical model is appropriate, the desirability of which is vividly displayed by Anscombe's quartet (e.g., Anscombe, 1973; Matejka & Fitzmaurice, 2017). Anscombe's quartet, shown in Figure 3, consists of four scatterplots; in all four, the summary statistics (i.e., sample size, mean, variance, and Pearson correlation) for the *x* and *y* variables are identical, and so are the Bayes factors; nevertheless, for three of the four scatterplots, an inference in terms of the strength of a linear association, based on the observed Pearson's correlation coefficient, is meaningless.[5]

We close by noting that the kind of Bayesian reanalyses outlined here provides an opportunity to expand summary statistics to statements about posterior distributions and Bayes factors. Such an expansion will afford (a) an additional inferential perspective to supplement the classical perspective, (b) reanalyses of published findings that do not require the raw data, and (c) a highly efficient method for editors, reviewers, readers, and reporters to gauge whether the conclusions from a different statistical paradigm contradict or confirm the conclusions obtained using classical methods. We hope that the Summary Stats module of JASP will spur more comprehensive, inclusive, and nuanced assessments of statistical evidence.

## Action Editor

Frederick L. Oswald served as action editor for this article.

## Author Contributions

E.-J. Wagenmakers generated the idea for the method. A. Ly and A. Raj implemented the module in JASP. A. Ly and

M. Marsman wrote the first draft of the manuscript, which was further developed by A. Ly and A. Etz, and all the authors critically edited it. Q. F. Gronau performed and drafted the section on informed analyses. All the authors approved the final submitted version of the manuscript.

## ORCID iDs

Alexander Etz https://orcid.org/0000-0001-9394-6804
Maarten Marsman https://orcid.org/0000-0001-5309-7502
Eric-Jan Wagenmakers https://orcid.org/0000-0003-1596-1034

## Acknowledgments

We thank Michael Masson, Sang Ho Lee, Jay Myung, and two reviewers for their constructive suggestions for improvement. We are also grateful to Raoul Grasman for his contributions to JASP.

## Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

## Funding

A. Ly and E.-J. Wagenmakers are supported by the starting grant “Bayes or Bust” awarded by the European Research Council (Grant 283876) and by Grant 016.Vici.170.083 from the Netherlands Organisation for Scientific Research (NWO). A. Etz was supported by Grant 1534472 from the National Science Foundation’s Methods, Measurements, and Statistics panel, as well as by the National Science Foundation Graduate Research Fellowship Program (Grant DGE1321846). Q. F. Gronau was supported by Grant 406.16.528 from NWO.

## Open Practices

The JASP file for the reanalyses discussed here (festingerCarlsmith1959.jasp) has been made publicly available via the Open Science Framework and can be accessed at https://osf.io/7t2jd/. The source code of JASP is publicly available at https://github.com/jasp-stats/jasp-desktop/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918779348. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

## Notes

[1] The calculators at https://web.archive.org/web/20170212075534/http://pcl.missouri.edu/bayesfactor, designed and maintained by Jeff Rouder, exploit the same idea, but are of more limited scope than the Summary Stats module.
[2] The Summary Stats module is activated via the “+” icon next to the “Common” tab at the top of the JASP window.
[3] Note that the prior distribution of the population effect size has shrunk the sample value of $d$ (i.e., 0.702) toward zero.
[4] For a further discussion of Festinger and Carlsmith’s (1959) results, see Heino (2016).
[5] See also Cairo’s (2016) Anscombosaurus.

## References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*, 17–21.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (2nd ed., Vol. 1, pp. 378–386). Hoboken, NJ: Wiley.

Cairo, A. (2016). Download the Datasaurus: Never trust summary statistics alone; always visualize your data [Web log post]. Retrieved from https://web.archive.org/web/20170901133148/http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

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

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

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, *25*, 219–234.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi:10.3758/s13423-017-1262-3

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane’s contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*, 203–210.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, *P* values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*, 337–350.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2017). Informed Bayesian *T*-tests. Retrieved from https://arxiv.org/pdf/1704.02479.pdf

Heino, M. (2016). The legacy of social psychology [Web log post]. Retrieved from https://web.archive.org/web/20170901132636/https://mattiheino.com/2016/11/13/legacy-of-psychology/

JASP Team. (2017). JASP (Version 0.8.2) [Computer software]. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford, England: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

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

Kelman, H. C. (1953). Attitude change as a function of response restriction. *Human Relations*, *6*, 185–214.

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (in press). Replication Bayes factors from evidence updating. *Behavior Research Methods*.

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55. doi:10.1016/j.jmp.2017.05.006

Ly, A., Verhagen, 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. doi:10.1016/j.jmp.2016.01.003

Ly, A., Verhagen, 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. doi:10.1016/j.jmp.2015.06.004

Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *CHI 2017 Conference Proceedings: ACM SIGCHI Conference on Human Factors in Computing Systems*. Retrieved from https://www.autodeskresearch.com/publications/samestats

Morey, R. D., & Rouder, J. N. (2015). BayesFactor (Version 0.9.11-1) [Computer software]. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, 25, 1–4.

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239.

Verhagen, 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., & Gronau, Q. F. (2015). *LetsPlayDarts.jpg (Version: 4)*. Retrieved from https://osf.io/3acm7/

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76. doi:10.3758/s13423-017-1323-7.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57. doi:10.3758/s13423-017-1343-3

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, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 123–138). Chichester, England: John Wiley & Sons. doi:10.1002/9781119095910.ch8

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

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.