Commentary

Psychological Science
2016, Vol. 27(7) 1043–1046

DOI: 10.1177/0956797616644081
pss.sagepub.com

Bayesian Evidence Synthesis Can Reconcile Seemingly Inconsistent Results: The Case of Hotel Towel Reuse

Benjamin Scheibehenne[1], Tahira Jamil[2], and Eric-Jan Wagenmakers[2]

[1]Faculty of Economics and Management, University of Geneva, and [2]Department of Psychological Methods, University of Amsterdam

Received 11/3/15; Revision accepted 3/21/16

Recent concerns that psychological science may suffer from a lack of replicability have prompted a methodological reorientation that values preregistration of hypotheses and data-analysis plans, high statistical power, exact replications, and the assessment of cumulative knowledge through meta-analysis (Eerland, Sherrill, Magliano, & Zwaan, 2016; Open Science Collaboration, 2015). This reorientation raises the question of how exactly new and old findings ought to be combined. Here, we outline a Bayesian approach that updates knowledge about an effect as new studies become available. This method—Bayesian evidence synthesis—affords several advantages: It provides a continuous measure of evidence that indexes the degree of support for the null hypothesis versus an alternative hypothesis (Monden et al., in press), it distinguishes between evidence for the absence of an effect versus absence of evidence for an effect (e.g., Dienes, 2014), and it allows a continual updating of knowledge as new studies appear, indefinitely and without a sampling plan or stopping rule (e.g., Rouder, 2014). Below, we highlight these advantages using a concrete example concerning the effectiveness of descriptive social norms in facilitating ecological behavior.

Descriptive social norms indicate which behavior is typical or normal in a given situation (Cialdini, Reno, & Kallgren, 1990). Such information can influence people's behavior in important ways (P. W. Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007). In a widely cited study on the effectiveness of descriptive social norms (Goldstein, Cialdini, & Griskevicius, 2008), two groups of hotel guests received different messages that encouraged them to reuse their towels. One message simply informed the guests about the benefits of environmental protection (the control condition), and the other message indicated that the majority of guests actually reused their towels in the past (the descriptive-social-norm condition). The results suggested that the latter message facilitated towel reuse (Experiment 1—descriptive-social-norm condition: 44.1% reuse, control condition: 35.1% reuse; $p = .05$; Experiment 2—descriptive-social-norm conditions (combined): 44.5% reuse, control condition: 37.2% reuse; $p = .03$).

A search across all studies in the literature that cited this original publication and a separate search combining the terms "social norm" and "towel reuse" revealed five replication experiments that assessed the proportion of hotel guests who reused their towels, with a total sample size of 2,466 participants (Bohner & Schlüter, 2014; Mair & Bergin-Seers, 2010; W. P. Schultz, Khazian, & Zaleski, 2008). All five experiments arguably failed to replicate the original finding (all ps > .14). However, this apparent contradiction can be resolved by a Bayesian reanalysis.

In the first step of this reanalysis,[1] we recorded how many participants reused their towel in each of the two conditions in all seven experiments. Next, for each experiment, we obtained a separate one-sided Bayes factor for a test of equality of two proportions (e.g., Gunel & Dickey, 1974; Jamil, Marsman, Ly, Morey, & Wagenmakers, in press; Jeffreys, 1961). In this analysis, the null hypothesis was that the proportions of guests who did and did not reuse their towels are equal, whereas the default alternative hypothesis was that the proportions are independent and uniformly distributed between 0 and 1, with the added restriction that the proportion in the

Corresponding Author:
Benjamin Scheibehenne, University of Geneva, Faculty of Economics and Management, 40 Boulevard du Pont d'Arve, CH-1211 Geneva, Switzerland
E-mail: benjamin.scheibehenne@unige.ch

descriptive-social-norm condition is higher than in the control condition.

Finally, we repeated the analysis for the combined data across all seven experiments. For each experiment individually and for the combined total, the posterior distribution of the log odds ratio and the corresponding Bayes factors in favor of the descriptive-social-norm hypothesis is displayed in the upper panel of Figure 1.[2]

The upper panel of Figure 1 reveals, first, that when considered in isolation, none of the experiments provides compelling evidence in favor of the descriptive-social-norm hypothesis. The strongest Bayes factor (BF) comes from Experiment 2 of Goldstein et al. (2008) and yields a modest BF_{10} of 2.03, which means that the data are only about twice as likely to be obtained if the alternative hypothesis is true than if the null hypothesis is true. Bayes factors below 3 are commonly considered ambiguous or anecdotal (Jeffreys, 1961). Three replication attempts also yielded only weak evidence. Thus, the apparent contradiction in the literature may be attributed in part to the infamous *p*-value "cliff effect" (Rosenthal & Gaito, 1963), that is, the tendency to believe that findings with *p*s less

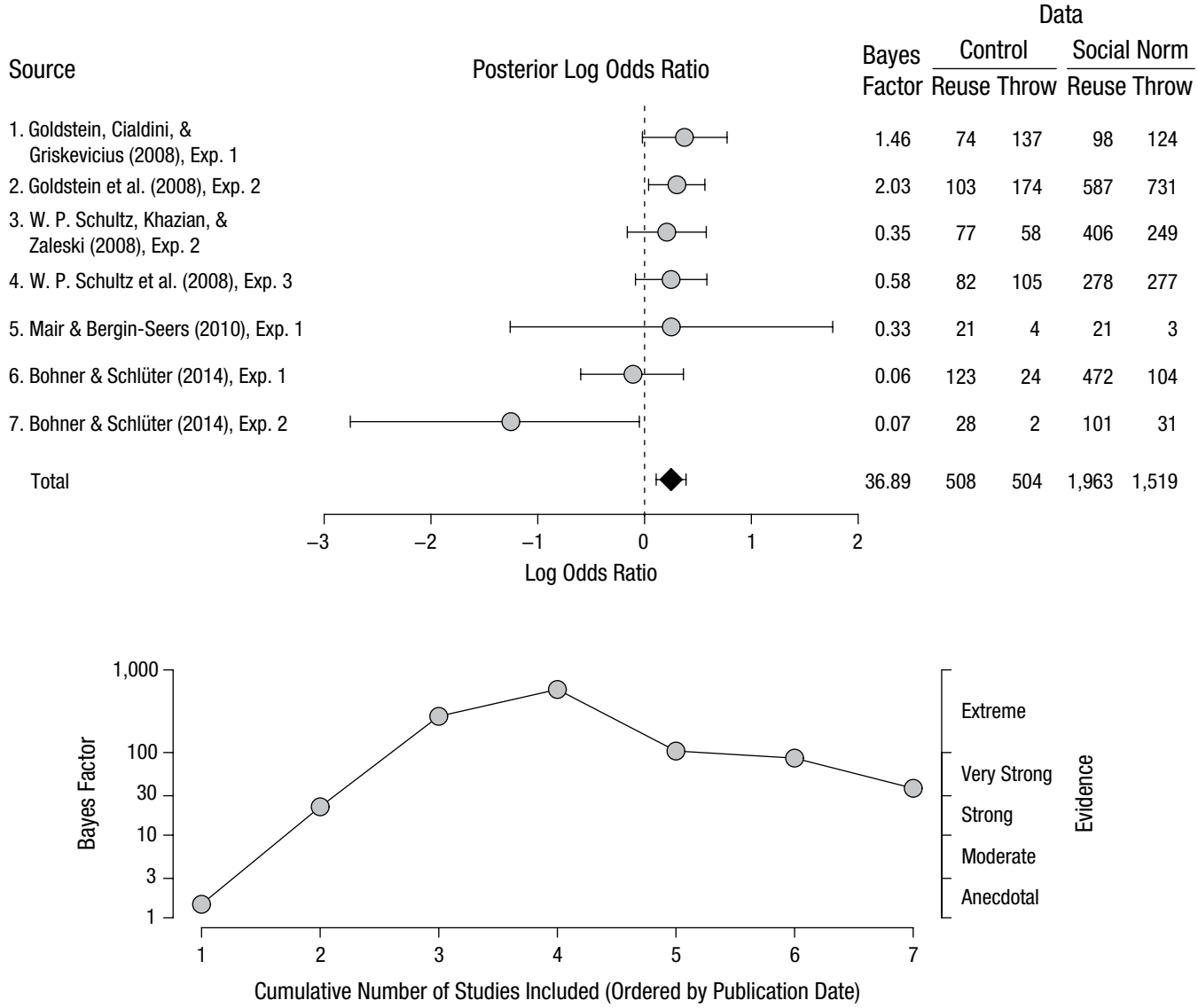

Fig. 1. Bayesian reanalysis of seven experiments on the effectiveness of social norms on reuse of hotel towels, separately for each experiment and ordered by publication date. In the upper panel, gray circles indicate the posterior mean of the log odds ratio, and error bars indicate 95% credible intervals. The corresponding Bayes factors in favor of the descriptive-social-norm hypothesis are also shown. Positive log odds ratios and Bayes factors greater than 1 indicate evidence for the effectiveness of social norms. Raw data are shown separately for participants who reused and for those who threw away their towels in each condition (descriptive social norm vs. control). The bottom panel shows the progression of the Bayes factor in favor of the descriptive-social-norm hypothesis as the experiments became available over time. This figure can also be downloaded from the Flickr Web site (https://flic.kr/p/E9XavM) and reproduced under Creative Commons license (https://creativecommons.org/licenses/by/2.0/).

than .05 are qualitatively different from those with *p*s greater than .05.

Figure 1 also reveals that, when collapsed across experiments, the data do provide strong evidence in favor of the descriptive-social-norm hypothesis: BF_{10} = 36.89, which means that the data are about 37 times more likely to be obtained if this hypothesis is true than if the null hypothesis is true. For the combined data, the mean log odds ratio is 0.25, and the 95% Bayesian credible interval ranges from 0.11 to 0.39. In absolute terms, the data indicate an average 6.2% increase in towel reuse in the descriptive-social-norm condition if the alternative hypothesis is true. Weighting this increase with the odds (i.e., the Bayes factor) in favor of the alternative hypothesis yields a model-averaged increase of 6.0% (6.2 × 36.89/(1 + 36.89)). This increase is modest, but—if veridical—its impact would nevertheless be substantial (e.g., in 2013, the European Union alone registered more than 1.7 billion overnight hotel bookings; Eurostat, 2015). Hotel owners who contemplate replacing their messages with ones suggesting that most guests reuse their towels may make an optimal decision by combining the evidence with an assessment of utilities (i.e., weighting the costs and benefits of the possible actions; Lindley, 1985).

Although other statistical models can be specified to analyze these findings (see the Supplemental Material available online), the qualitative patterns of results is robust: When analyzed individually, none of the experiments provides compelling evidence for the effectiveness of descriptive social norms on towel reuse, but together the experiments provide strong support. Note that the analysis on the combined data assumed a fixed effect; with only seven experiments and in the absence of strong prior knowledge, we feel that a random-effects analysis would be overly ambitious. Also, as our analysis included only published data, its results are not immune to file-drawer problems or publication biases. Nevertheless, our results do suggest a new interpretation of the nonsignificant findings.

As an alternative to Bayesian evidence synthesis, a classical fixed-effects meta-analysis yields a mean log odds ratio of 0.23 (95% confidence interval = [0.072, 0.381], p = .004) and is numerically consistent with the Bayesian results. However, the classical analysis is unable to quantify evidence for the null hypothesis, and it cannot discriminate between evidence for absence of an effect versus absence of evidence for an effect (Dienes, 2014). Moreover, the classical results—*p* values as well as confidence intervals—require adjustment depending on the sampling plan and the stopping rule (Berger & Wolpert, 1988). When studies arrive sequentially without a well-defined stopping rule, the sample space of possible outcomes is ill-defined, and the correction for multiple testing becomes problematic (for classical sequential meta-analysis, see Higgins, Whitehead, & Simmonds, 2011; Wetterslev, Thorlund, Brok, & Gluud, 2008). In contrast, the Bayesian approach relies on the data that were actually observed and allows evidence to be seamlessly updated after every new study. As an example, the lower panel of Figure 1 shows the evidential trajectory for the descriptive-social-norm-hypothesis over time. Also, if additional data on towel reuse become available, the analysis can be updated easily.

In sum, Bayesian evidence synthesis is a promising meta-analytic approach. The example of towel reuse demonstrated how adding new studies lengthens the evidential trajectory (Fig. 1, lower panel). Furthermore, our Bayesian analysis revealed that all current findings on towel reuse are evidentially weak; when analyzed together, however, they provide support for the hypothesis that descriptive social norms can prompt people to alter their behavior in ways that may benefit the environment.

Action Editor

D. Stephen Lindsay served as action editor for this article.

Author Contributions

B. Scheibehenne collected the data. All authors contributed to the analysis of the data. B. Scheibehenne and E.-J. Wagenmakers wrote the manuscript. All authors approved the final version of the manuscript for submission.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by Swiss National Science Foundation Grant No. 1000014_149846 to the first author and by European Research Council Grant No. 283876 to the third author.

Supplemental Material

Additional supporting information can be found at http://pss.sagepub.com/content/by/supplemental-data

Open Practices

All data and materials have been made publicly available via the Open Science Framework and can be accessed at https://osf.io/ycx49/. The complete Open Practices Disclosure for this article can be found at http://pss.sagepub.com/content/by/supplemental-data. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ and http://pss.sagepub.com/content/25/1/3.full.

Notes

1. A detailed tutorial-style description of this reanalysis is in the Supplemental Material available online and also at Open Science Framework (https://osf.io/tz6xv/).
2. The Supplemental Material also contains the raw data and the R code underlying Figure 1.

References

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

Bohner, G., & Schlüter, L. E. (2014). A room with a viewpoint revisited: Descriptive norms and hotel guests' towel reuse behavior. *PLoS ONE*, *9*(8), Article e106606. doi:10.1371/journal.pone.0106606

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*, 1015–1026.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, Article 781. doi:10.3389/fpsyg.2014.00781

Eerland, A., Sherrill, A. M., Magliano, J. P., & Zwaan, R. A. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science*, *11*, 158–171.

Eurostat. (2015). *Eurostat regional yearbook 2015*. Retrieved from http://ec.europa.eu/eurostat/publications/regional-yearbook

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, *35*, 472–482.

Gunel, E., & Dickey, J. (1974). Bayes factors for independence in contingency tables. *Biometrika*, *61*, 545–557.

Higgins, J. P. T., Whitehead, A., & Simmonds, M. (2011). Sequential methods for random-effects meta-analysis. *Statistics in Medicine*, *30*, 903–921.

Jamil, T., Marsman, M., Ly, A., Morey, R. D., & Wagenmakers, E.-J. (in press). What are the odds? Modern relevance and Bayes factor solutions for MacAlister's problem from the 1881 *Educational Times*. *Educational and Psychological Measurement*.

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

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

Mair, J., & Bergin-Seers, S. (2010). The effect of interventions on the environmental behaviour of Australian motel guests. *Tourism and Hospitality Research*, *10*, 255–268.

Monden, R., de Vos, S., Morey, R. D., Wagenmakers, E.-J., de Jonge, P., & Roest, A. M. (in press). Toward evidence-based medical statistics: A Bayesian analysis of double-blind placebo-controlled antidepressant trials in the treatment of anxiety disorders. *International Journal of Methods in Psychiatric Research*.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *The Journal of Psychology*, *55*, 33–38.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*, 301–308.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, *18*, 429–434.

Schultz, W. P., Khazian, A. M., & Zaleski, A. C. (2008). Using normative social influence to promote conservation among hotel guests. *Social Influence*, *3*, 4–23.

Wetterslev, J., Thorlund, K., Brok, J., & Gluud, C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology*, *61*, 64–75.