Psychological Methods
2017, Vol. 22, No. 2, 211–216

© 2017 American Psychological Association
1082-989X/17/$12.00 http://dx.doi.org/10.1037/met0000143

INTRODUCTION

Bayesian Hypothesis Testing: Editorial to the Special Issue on Bayesian Data Analysis

Herbert Hoijtink
Utrecht University

Sy-Miin Chow
Pennsylvania State University

In the past 20 years, there has been a steadily increasing attention and demand for Bayesian data analysis across multiple scientific disciplines, including psychology. Bayesian methods and the related Markov chain Monte Carlo sampling techniques offered renewed ways of handling old and challenging new problems that may be difficult or impossible to handle using classical approaches. Yet, such opportunities and potential improvements have not been sufficiently explored and investigated. This is 1 of 2 special issues in *Psychological Methods* dedicated to the topic of Bayesian data analysis, with an emphasis on Bayesian hypothesis testing, model comparison, and general guidelines for applications in psychology. In this editorial, we provide an overview of the use of Bayesian methods in psychological research and a brief history of the Bayes factor and the posterior predictive *p* value. Translational abstracts that summarize the articles in this issue in very clear and understandable terms are included in the Appendix.

Keywords: Bayes Factor, Bayesian data analysis, Bayesian estimation, Bayesian modeling, posterior predictive *p* value

In 2015, *Psychological Methods* published a call for articles for a special issue on Bayesian Data Analysis. A total of 65 articles have been submitted of which approximately 20 will be published. The latter can be categorized under five topics: practical guidelines and general use, Bayes factor, posterior predictive *p* values, Bayesian estimation, and Bayesian modeling. We decided to distribute these articles over two special issues: the June 2017 issue containing the articles with respect to the first three topics and the December 2017 issue containing the articles with respect to the last two topics. As much as possible, we have encouraged authors to provide software codes and practical demonstrations with their articles. These are available online as supplementary materials.

This editorial and special issue are organized into three sections. A summary of the articles included is presented, followed by the Appendix, which provides translational abstracts of the articles briefly describing their essence in clear understandable language. We begin with an overview of Bayesian data analysis in psychological research, thereby introducing the articles by van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017) and Depaoli and van de Schoot (2017). The editorial continues with a brief history of the Bayes factor and its use in psychological research, thereby introducing the articles by Böing-Messing, van Assen, Hofman, Hoijtink, and Mulder (2017); Rouder, Morey, Verhagen, Swagman, and Wagenmakers (2017); Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017); Houpt, Heathcote, and Eidels (2017); Jeon and De Boeck (2017), and Lu, Chow, and Loken (2017). Subsequently, a brief history of the posterior predictive *p* value and its use in psychological research will be provided, thereby introducing the articles by van Kollenburg, Mulder, and Vermunt (2017) and Li, Xie, and Jiao (2017). The editorial ends with a short conclusion.

Overview of Bayesian Data Analysis in Psychological Research

Most modern applications of Bayesian data analysis in psychological research employ a computationally demanding Gibbs or Markov chain Monte Carlo sampler (McMC; Gelfand & Smith, 1990; Geman & Geman, 1984). The speed of computers has always been an important determinant—and hurdle, in many cases—of the rates at which Bayesian data analysis develops and gains momentum in psychology. In February 1994, the first author of this editorial programmed in Fortran an McMC approach to sample the parameters of the two-parameter logistic item response model (Birnbaum, 1968) on a, for that time, state-of-the-art laptop. After 2 weeks of sampling, the program finished. Inspection of the output showed that there was a bug in the code. To make a long story short, some initial results were only available toward the end

Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University; Sy-Miin Chow, Department of Human Development and Family Studies, Pennsylvania State University.

The first author is supported by the Consortium on Individual Development (CID), which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant 024.001.003). Funding for the second author was provided by NSF grant SES-1357666 and NIH grant R01GM105004. We are indebted to the editor of *Psychological Methods*, Lisa Harlow, for the opportunity to oversee this special issue and for her support and feedback throughout the editorial process. There has been no prior dissemination of the ideas expressed in this article before it was submitted for publication to *Psychological Methods*.

Correspondence concerning this article should be addressed to Herbert Hoijtink, Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, the Netherlands. E-mail: H.Hoijtink@uu.nl

of April 1994. In 2015, the exercise was repeated using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter 2000), again on a (for that time) state-of-the-art computer. This time, computations finished in a few minutes. Aside from the anecdotal coincidence with the first author's personal experience, the time frame (1990–2015) for van de Schoot and colleagues' (2017, this issue) review of Bayesian data analysis in psychological research is, indeed, well-chosen. Before 1990, there were virtually no "psychological" articles dealing with applied Bayesian statistics (but see Edwards, Lindman, & Savage, 1963); however, with increases in computing resources and available software, the number of publications on this topic has rapidly multiplied.

The Bayesian point of view and computational toolkit enable new approaches to data analysis. However, as is highlighted by Depaoli and van de Schoot (2017, this issue), new approaches also call for new standards and guidelines for reporting analyses to improve transparency and replicability across studies. Their WAMBS checklist (*W*hen to worry and how to *A*void the *M*isuse of *B*ayesian *S*tatistics) provides one possible set of recommendations on important results to report. Their checklist is timely, and we agree with the authors that such a checklist provides a working framework to help researchers organize their analytic plans and results. Two important new approaches are testing hypotheses using the Bayes factor and the posterior predictive *p* value. In the next two sections, both approaches will be introduced.

Brief History of the Bayes Factor and Its Use in Psychological Research

The core of the Bayesian approach is Bayes's theorem, which states that the posterior density $g(\cdot)$ is the product of the density of the data $f(\cdot)$ and the prior distribution $h(\cdot)$ divided by the marginal likelihood $m(\cdot)$:

$$g(\theta | x) = \frac{f(x | \theta) h(\theta)}{m(x)}, \tag{1}$$

where x denotes the data, and θ the parameters of the model defining the density of the data. In classical statistics, the information in the data with respect to the model parameters is summarized in the likelihood function. In Bayesian statistics, this information is summarized in the posterior distribution. A distinguishing feature of the Bayesian approach is the opportunity to include prior knowledge about the model parameters into statistical analyses via the specification of prior distributions. The interested reader is referred to Gelman, Carlin, Stern, and Rubin (2013) for an introduction and many examples. Another feature is the use of the marginal likelihood, which contains the information in the data with respect to a model, to test hypotheses. Bayesian hypothesis testing by means of the Bayes factor was introduced by Jeffreys (1939/1961). In psychological research, it is often used in the form of the following ratio of two marginal likelihoods:

$$BF_{12} = \frac{m(x | H1)}{m(x | H2)} = \frac{\int_\theta f(x | \theta) h(\theta | H_1) d\theta}{\int_\theta f(x | \theta) h(\theta | H_2) d\theta}, \tag{2}$$

where the marginal likelihoods and prior distributions are now conditional on H_1 and H_2, respectively, where "H" refers to a specific hypothesis. If, for example, $BF_{12} = 10$, the support in the data for H_1 is 10 times larger than the support for H_2. The density of the data can, for example, be based on the simple model $x_i \sim \mathcal{N}(\mu, \sigma^2)$ for $I = 1, \ldots, N$, where μ and σ^2 denote the population mean and variance of x, respectively, and N the sample size. To obtain a Bayesian test of the hypotheses $H_1 : \mu = 0$ versus $H_2 : \mu \neq 0$, the prior distributions have to be specified. Prior distributions could be, for example, $h(\mu | H_1) = I_{\mu=0}$, a point-mass or probability of 1 for $\mu = 0$, $h(\mu | H_2) = \mathcal{N}(0, \tau)$, where τ is the prior variance; and a standard uninformative prior for σ^2.

An important landmark reviving the attention for the Bayes factor is Kass and Raftery (1995). Readers wanting to learn about the Bayes factor are well advised to start with this article. The main question when using Bayes factors is how to specify the prior distributions. In our example this would amount to choosing a value for τ. Currently, within psychological research two main approaches can be distinguished. The first approach uses data based methods to specify the prior distribution (see Berger and Pericchi, 2004, for an overview). The first appearance of the Bayes factor in *Psychological Methods*, Klugkist et al. (2005), used data based prior distributions to evaluate so-called informative hypotheses; that is, hypotheses specified using equality and inequality constraints (Hoijtink, 2012). The interested reader is referred to http://informative-hypotheses.sites.uu.nl/for an overview and software packages with which informative hypotheses can be evaluated. The latest extension of this approach is Böing-Messing et al. (2017, this issue), who evaluate inequality constrained hypotheses with respect to variances. An application of this approach can be found in Houpt et al. (2017, this issue) who test informative hypotheses with respect to cognitive architecture. The second approach uses (mixtures of) so-called g-priors (Liang, Paulo, Molina, Clyde, & Berger 2008). In our example these priors would be based on a subjective choice of the value of τ. Rouder, Speckman, Sun, Morey, and Iverson (2009) is the starting point of a number of g-prior based Bayes factor articles for the evaluation of traditional null-hypotheses. The interested reader is referred to http://pcl.missouri.edu/bayesfactor for an overview and software packages in which this approach is implemented. The latest extension is Rouder et al. (2017, this issue) and concerns Bayesian analysis of factorial designs. An application of this approach can be found in Schönbrodt et al. (2017, this issue) who use it in the context of sequential hypothesis testing. Both approaches are (being) implemented in the software package JASP (https://jasp-stats.org/) thereby increasing their usability for psychological researchers.

Bayes factor is not the only mechanism for model comparison. It has its strengths in some modeling contexts, but also its limitations in others. Comparisons of the Bayes factor to other model comparison tools are undertaken in several articles that appear in this special issue. Among them are the article by Jeon and De Boeck (2017, this issue), in which they compare the decision qualities of the Bayes factor and *p* value based hypothesis testing, and that by Lu et al. (2017, this issue) in which the authors compare the performance of the Bayes factor with model comparison criteria such as the Bayesian Information Criterion and the Deviance Information Criterion, as well as alternative approaches based on Bayesian leave-one-out and variable selection methods.

A Brief History of the Posterior Predictive *p* Value and Its Use in Psychological Research

The first proposal for the posterior predictive *p* value was put forth by Rubin (1984, Section 5) as a mechanism for testing a null hypothesis, H_0, with respect to a model of interest. We use the name "null hypothesis" for convenience, but generally, H_0 may take on a form that is much more general than the way the null hypothesis is typically defined in the classical sense. If the model of interest is, for example, $x_i \sim \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, N$, the hypothesis could be (a) $H_0^a{:}\mu = 0$ and (b) $H_0^b{:}x_i$ is normally distributed. The formal definition of the posterior predictive p value is:

$$p = P(T_{rep} > T|x,\ H_0) = \int_{\theta_0} P(T_{rep} > T|\theta_0) g(\theta_0|x,\ H_0) d\theta_0$$
$$\approx \frac{1}{Q}\sum_{q=1}^{Q} I_{T_{rep}^q > T}, \quad (3)$$

where the vector of parameters for the model defined under H_0 is denoted as θ_0. For the two examples listed above, $\theta_0 = [\mu = 0, \sigma^2]$ and $[\mu, \sigma^2]$, respectively, under H_0^a and H_0^b. The posterior distribution of θ_0 is denoted by $g(\theta_0 \,|\, x, H_0)$, which is proportional to $f(x \,|\, \theta_0)h(\theta_0)$, where $h(\cdot)$ represents the prior distributions for θ_0, usually selected to be uninformative in nature. The plausibility of H_0 is assessed using T, a sample statistic of choice computed using the observed data. For example, for H_0^a, $T = \dfrac{\bar{x}}{s/\sqrt{N}}$, where $\bar{x}$ and s^2 denote the sample mean and variance of x, respectively. In the case of H_0^b, choices for T may include the sample skewness, the sample kurtosis, or the largest value of x.

A McMC sampling method can be used to draw $q = 1, \ldots Q$ samples of θ_0^q from $g(\cdot)$. Each θ_0^q is, in turn, used to generate x_{rep}^q, a replication of the data matrix sampled from the posterior predictive distribution of the data under H_0. Each x_{rep}^q can be used to compute T_{rep}^q, thereby rendering a sample from the posterior predictive distribution of the test statistic under H_0. Simply counting the proportion of times that the replicated test statistics are larger than the observed test statistic provides an estimate of the posterior predictive p value.

The posterior predictive approach provides one possible solution to the fundamental problem of computing p values: how to replicate data from a population in which H_0 is true if one or more of the population parameters are unknown. In the case of H_0^a, only σ^2 is unknown; in the case of H_0^b, both μ and σ^2 are unknown. The problem is addressed by sampling the unknown parameters from their joint posterior distribution using one of a variety of McMC sampling techniques.

An important landmark in the development of the posterior predictive p value is Meng (1994). He provided a formal description, elaborations, examples, and derived properties. As became clear, the posterior predictive p value is not necessarily uniform under the null-hypothesis; that is, it may very well not hold that $P(p < \alpha \,|\, H_0) = \alpha$, where α denotes the Type I error level. Bayarri and Berger (2000) elaborated that this bias is caused by using the data twice: once for the specification of the posterior distribution and once for the computation of the p value. They also provide modifications that solve this problem; however, these modifications are not easily applied. Van Kollenburg et al. (2017, this issue) provide another solution (posterior calibration of the p values) and apply it in the context of latent class models and regression analysis. The interested reader is also referred to Lecoutre, Lecoutre, and Poitevineau (2010), who discuss the use of predictive probabilities in psychological research and constitute the first appearance of predictive probabilities in *Psychological Methods*.

Despite its nonuniformity, proponents of the posterior predictive p value argue that it is very useful because it is a clearly defined and easily applied model check (Gelman, Meng, & Stern, 1996) and can always be used as such. Li et al. (2017, this issue) show how model checking using the posterior predictive p value can be used to assess the fit of unidimensional polytomous item response theory (IRT) models. It is furthermore implemented in packages such as Mplus (Muthén & Muthén, 1998–2015) and Blavaan (Merkle & Rosseel, 2015), where it is used to test the fit of structural equation models.

Conclusion

As van de Schoot et al. (2017, this issue) show, the history of Bayesian data analysis in psychological research started about 25 years ago. Since then, many new developments, applications, and refinements to existing techniques have been achieved. We have highlighted here only a fraction of the new developments in theory and software packages. We do not claim the collection of articles appearing in this special issue to be exhaustive. Nevertheless, we hope that these articles can add to the repertoire of Bayesian tools and resources available to psychological researchers in useful ways.

References

Bayarri, M. J., & Berger, J. O. (2000). *p* Values for composite null models. *Journal of the American Statistical Association, 95,* 1127–1142. http://dx.doi.org/10.1080/01621459.2000.10474309

Berger, J. O., & Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *The Annals of Statistics, 32,* 841–869. http://dx.doi.org/10.1214/009053604000000229

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, PA: Addison Wesley.

Böing-Messing, F., van Assen, M. A. L. M., Hofman, A. D., Hoijtink, H., & Mulder, J. (2017). Bayesian evaluation of constrained hypotheses on variances of multiple independent groups. *Psychological Methods, 22,* 262–287. http://dx.doi.org/10.1037/met0000116

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22,* 240–261. http://dx.doi.org/10.1037/met0000065

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242. http://dx.doi.org/10.1037/h0044139

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409. http://dx.doi.org/10.1080/01621459.1990.10476213

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistics Sinica, 6,* 733–807.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Patterns Analysis and Machine Intelligence, 6,* 721–741. http://dx.doi.org/10.1109/TPAMI.1984.4767596

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. Boca Raton, FL: Chapman and Hall/CRC.

Houpt, J. W., Heathcote, A., & Eidels, A. (2017). Bayesian analyses of cognitive architecture. *Psychological Methods, 22,* 288–303. http://dx.doi.org/10.1037/met0000117

Jeffreys, H. (1939/1961). *Theory of probability* (1st/3rd ed.). Oxford, United Kingdom: Oxford University.

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p-value based hypothesis testing. *Psychological Methods, 22,* 340–360. http://dx.doi.org/10.1037/met0000140

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10,* 477–493. http://dx.doi.org/10.1037/1082-989X.10.4.477

Lecoutre, B., Lecoutre, M.-P., & Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use, and interpret them. *Psychological Methods, 15,* 158–171. http://dx.doi.org/10.1037/a0015915

Li, T., Xie, C., & Jiao, H. (2017). Assessing fit of alternative unidimensional polytomous IRT models using posterior predictive model checking. *Psychological Methods, 22,* 397–408. http://dx.doi.org/10.1037/met0000082

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103,* 410–423. http://dx.doi.org/10.1198/016214507000001337

Lu, Z.-H., Chow, S.-M., & Loken, E. (2017). A comparison of Bayesian and frequentist model selection methods for factor analysis models. *Psychological Methods, 22,* 361–381. http://dx.doi.org/10.1037/met0000145

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10,* 325–337. http://dx.doi.org/10.1023/A:1008929526011

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics, 22,* 1142–1160. http://dx.doi.org/10.1214/aos/1176325622

Merkle, E. C., & Rosseel, Y. (2015). *blavaan: Bayesian structural equation models via parameter expansion*. Retrieved from http://arxiv.org/abs/1511.05604

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author. Retrieved from: https://www.statmodel.com/

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods, 22,* 304–321. http://dx.doi.org/10.1037/met0000057

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 and Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics, 12,* 1151–1172. http://dx.doi.org/10.1214/aos/1176346785

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22,* 322–339. http://dx.doi.org/10.1037/met0000061

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods, 22,* 217–239. http://dx.doi.org/10.1037/met0000100

van Kollenburg, T., Mulder, J., & Vermunt, J. (2017). Posterior calibration of posterior predictive p-values, with application in latent class and regression analysis. *Psychological Methods, 22,* 382–396. http://dx.doi.org/10.1037/met0000142

Appendix

Translational Abstracts (TAs) for the 10 Special Issue Articles

1. TA for "A Systematic Review of Bayesian Papers in Psychology: The Last 25 Years" by Rens van de Schoot, Sonja Winter, Oisin Ryan, Marielle Zondervan-Zwijnenburg, and Sarah Depaoli

Although the statistical tools most often used by researchers in the field of Psychology over the last 25 years are based on frequentist statistics, it is often claimed that the alternative Bayesian approach to statistics is gaining in popularity. In the current article, we investigated this claim by performing the very first systematic review of Bayesian psychological articles published between 1990 and 2015 (n = 1,579). We aim to provide a thorough presentation of the role Bayesian statistics plays in Psychology. This historical assessment allows us to identify trends and see how Bayesian methods have been integrated into psychological research in the context of different statistical frameworks (e.g., hypothesis testing, cognitive models, IRT, structural equation modeling, etc.). We also describe take-home messages and provide "big-picture" recommendations to the field as Bayesian statistics becomes more popular. Our review indicated that Bayesian statistics are used in a variety of contexts across subfields of Psychology and related disciplines. There are many different reasons why one might choose to use Bayes (e.g., the use of priors, estimating otherwise intractable models, modeling uncertainty, etc.). We found in this review that the use of Bayes has increased and broadened in the sense that this methodology can be used in a flexible manner to tackle many different forms of questions. We hope this presentation opens the door for a larger discussion regarding the current state of Bayesian statistics, as well as future trends.

2. TA for "Improving Transparency and Replication in Bayesian Statistics: The WAMBS Checklist" by Sarah Depaoli and Rens van de Schoot

Bayesian statistical methods are slowly creeping into all fields of science and are becoming ever more popular in applied research. Although it is very attractive to use Bayesian statistics, our personal experience has led us to believe that naively applying Bayesian methods can be dangerous for at least three main reasons: the potential

(*Appendix continues*)

influence of priors, misinterpretation of Bayesian features and results, and improper reporting of Bayesian results. To deal with these three points of potential danger, we have developed a succinct checklist: the WAMBS-checklist (When to Worry and how to Avoid the Misuse of Bayesian Statistics). The purpose of the questionnaire is to describe 10 main points that should be thoroughly checked when applying Bayesian analysis. We provide an account of ”when to worry” for each of these issues related to: (a) issues to check before estimating the model, (b) issues to check after estimating the model but before interpreting results, (c) understanding the influence of priors, and (d) actions to take after interpreting results. To accompany these key points of concern, we will present diagnostic tools that can be used in conjunction with the development and assessment of a Bayesian model. We also include examples of how to interpret results when “problems” in estimation arise, as well as syntax and instructions for implementation. Our aim is to stress the importance of openness and transparency of all aspects of Bayesian estimation, and it is our hope that the WAMBS questionnaire can aid in this process.

3. TA for “Bayesian Evaluation of Constrained Hypotheses on Variances of Multiple Independent Groups” by Florian Böing-Messing, Marcel A.L.M. van Assen, Abe D. Hofman, Herbert Hoijtink, and Joris Mulder

Research has shown that independent groups often differ not only in their means, but also in their variances. Comparing and testing variances is therefore of crucial importance to understand the effect of a grouping variable on an outcome variable. Researchers may have specific expectations concerning the relations between the variances of multiple groups. Such expectations can be translated into hypotheses with inequality and/or equality constraints on the group variances. Currently, however, no methods are available for testing (in)equality constrained hypotheses on variances. This article proposes a novel Bayesian approach to this challenging testing problem. Our approach has the following useful properties: First, it can be used to simultaneously test multiple (non)nested hypotheses with equality as well as inequality constraints on the variances. Second, our approach is fully automatic in the sense that no subjective prior specification is needed. Only the hypotheses need to be provided. Third, a user-friendly software application is included that can be used to perform this Bayesian test in an easy manner.

4. TA for “Bayesian Analyses of Cognitive Architecture” By Joseph W. Houpt, Andrew Heathcote, and Ami Eidels

The question of cognitive architecture—how cognitive processes are temporally organized—has arisen in many areas of psychology. This question has proved difficult to answer, with many proposed solutions turning out to be spurious. Systems Factorial Technology provided the first rigorous empirical and analytical method of identifying cognitive architecture, using the Survivor Interaction Contrast (SIC) to determine when people are using multiple sources of information in parallel or in series. Although the SIC is based on rigorous nonparametric mathematical modeling of response time distributions, for many years inference about cognitive architecture has relied solely on visual assessment. Recently, null hypothesis significance tests were introduced, and here we develop both parametric and nonparametric (encompassing prior) Bayesian inference. We show that the Bayesian approaches can have considerable advantages.

5. TA for “Bayesian Analysis of Factorial Designs” by Jeffrey N. Rouder, Richard D. Morey, Josine Verhagen, April R. Swagman, and Eric-Jan Wagenmakers

This article provides a Bayes factor approach to multiway analysis of variance (ANOVA) that allows researchers to state graded evidence for effects or invariances as determined by the data. ANOVA is conceptualized as a hierarchical model where levels are clustered within factors. The development is comprehensive in that it includes Bayes factors for fixed and random effects and for within-subjects, between-subjects, and mixed designs. Different model construction and comparison strategies are discussed, and an example is provided. We show how Bayes factors may be computed with BayesFactor package in R and with the JASP statistical package.

6. TA for “Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences” by Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini

Unplanned optional stopping rules have been criticized for inflating Type I error rates under the null hypothesis significance testing (NHST) paradigm. Despite these criticisms this research practice is not uncommon, probably as it appeals to researcher’s intuition to collect more data in order to push an indecisive result into a decisive region. In this contribution we investigate the properties of a procedure for Bayesian hypothesis testing that allows optional stopping with unlimited multiple testing, even after each participant. In this procedure, which we call Sequential Bayes Factors (SBF), Bayes factors are computed until an a priori defined level of evidence is reached. This allows flexible sampling plans and is not dependent upon correct effect size guesses in an a priori power analysis. We investigated the long-term rate of misleading evidence, the average expected sample sizes, and the biasedness of effect size estimates when an SBF design is applied to a test of mean differences between two groups. Compared with optimal NHST, the SBF design typically needs 50% to 70% smaller samples to reach a conclusion about the presence of an effect, while having the same or lower long-term rate of wrong inference.

(*Appendix continues*)

7. TA for "Decision qualities of Bayes factor and p-value based hypothesis testing" by Minjeong Jeon and Paul De Boeck

The purpose of this article is to investigate the decision qualities of the Bayes factor method compared with the *p* value based null hypothesis significance testing (NHST). The performance of the two methods is assessed in terms of the false and true positive rates as well as the false discovery rates and the posterior probabilities of the null hypothesis for two different models: an independent-samples t-test and an ANOVA model with two random factors. Our simulation study results showed the following: (1) The common Bayes factor $>$ 3 criterion is more conservative than the NHST alpha = .05 criterion, and it corresponds better with the alpha = .01 criterion. (2) An increasing sample size has a different effect on the false positive rate and the false discovery rate depending on whether the Bayes factor or NHST approach is used. (3) When effect sizes are randomly sampled from the prior, power curves tend to be flat compared with when effect sizes are prespecified. (4) The larger the scale factor (or the wider the prior) is, the more conservative the inferential decision is. (5) The false-positive and true-positive rates of the Bayes factor method are very sensitive to the scale factor when the effect size is small. (6) While the posterior probabilities of the null hypothesis ideally follow from the BF value, they can be surprisingly high using NHST. In general, these findings were consistent independent of which of the two different models was used.

8. TA for "A Comparison of Bayesian and Frequentist Model Selection Methods for Factor Analysis Models" by Zhaohua Lu, Sy-Miin Chow, and Eric Loken

We compare the performances of well-known frequentist model fit indices (MFIs) and several Bayesian model selection criteria (MSC) as tools for cross-loading selection in factor analysis under low to moderate sample sizes, effect sizes, and possible violation of distributional assumptions. The Bayesian criteria considered include the Bayes factor (BF), Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), Bayesian leave-one-out approach based on Pareto-smoothed importance sampling (LOO-PSIS), and a Bayesian variable selection method using the spike-and-slab prior (SSP; Lu et al., 2017). Simulation results indicate that the BF and the SSP showed the best balance between true positive rates and false positive rates, followed closely by the BIC. The SSP actually exhibited better performance than the BF as computed using the bridge sampler. The LOO-PSIS and the DIC showed the highest true positive rates among all the measures considered, but both had elevated false positive rates. In comparison, likelihood ratio tests (LRTs) are still the preferred frequentist model comparison tool, showing comparable or even higher true positive rates than the BF, SSP and BIC; under violations of distributional assumptions, however, slightly higher false positive detection rates were observed than for the Bayesian MCC. The root mean squared error of approximation (RMSEA), at the conventional cut-off of approximate fit, imposes a much more stringent "penalty" under conditions with low effect size, low sample size, and high model complexity compared with the LRTs and all other Bayesian MCC. Nevertheless, it provided a reasonable alternative to the LRTs in cases where the models cannot be constructed as nested within each other.

9. TA for "Posterior Calibration of Posterior Predictive P-values, with Applications in Latent Class and Regression Analysis" by Geert Hein van Kollenburg, Joris Mulder, and Jeroen K. Vermunt

To accurately control the type I error probability (typically.05), a *p* value should be uniformly distributed under the null model. The posterior predictive *p* value (ppp), which is commonly used in Bayesian data analysis, generally does not satisfy this property. For example there have been reports where the sampling distribution of the ppp under the null model was highly concentrated around .50. In this case, a ppp of.20 would indicate model misfit, but when comparing it with a significance level of.05, which is standard statistical practice, the null model would not be rejected. Therefore, the ppp has very little power to detect model misfit. A solution has been proposed in the literature, which involves calibrating the ppp using the prior distribution of the parameters under the null model. A disadvantage of this method is, however, that it is very sensitive to the quality of prior information that is provided about all model parameters. In this article, therefore an alternative solution is proposed where the ppp is calibrated using the posterior under the null model. This method (a) can be used when good prior information is absent, (b) allows one to test any model assumption by choosing an appropriate discrepancy measure, and (c) results in *p* values that are uniformly distributed under the null model. The new methodology is applied in various testing problems such as assessing model misfit in latent class analysis and checking misfit with outliers in linear regression.

10. TA for "Assessing Fit of Alternative Unidimensional Polytomous IRT Models Using Posterior Predictive Model Checking" by Tongyun Li, Chao Xie, and Hong Jiao

This article explored the application of a posterior predictive model checking (PPMC) method in assessing fit for unidimensional polytomous item response theory (IRT) models, specifically the divide-by-total models (e.g., the generalized partial credit model). Previous research has primarily focused on using PPMC in model checking for unidimensional and multidimensional IRT models for dichotomous data and paid little attention to polytomous models. A Monte Carlo simulation was conducted to investigate the performance of PPMC in detecting different sources of misfit for the partial credit model family. Results showed that the PPMC method, in combination with appropriate discrepancy measures, had adequate power in detecting different sources of misfit for the partial credit model family. Global odds ratio and item total correlation exhibited a specific pattern in detecting the absence of the slope parameter, whereas another method, Yen's Q1, was found to be promising in the detection of misfit caused by the constant category intersection parameter constraint across items.

Received January 20, 2017
Revision received February 6, 2017
Accepted February 8, 2017 ■