European Journal of Developmental Psychology

ISSN: 1740-5629 (Print) 1740-5610 (Online) Journal homepage: https://www.tandfonline.com/loi/pedp20

An introduction to Bayesian model selection for evaluating informative hypotheses

Rens van de Schoot , Joris Mulder , Herbert Hoijtink , Marcel A. G. Van Aken , Judith Semon Dubas , Bram Orobio de Castro , Wim Meeus & Jan-Willem Romeijn

To cite this article: Rens van de Schoot , Joris Mulder , Herbert Hoijtink , Marcel A. G. Van Aken , Judith Semon Dubas , Bram Orobio de Castro , Wim Meeus & Jan-Willem Romeijn (2011) An introduction to Bayesian model selection for evaluating informative hypotheses, European Journal of Developmental Psychology, 8:6, 713-729, DOI: 10.1080/17405629.2011.621799

To link to this article: https://doi.org/10.1080/17405629.2011.621799

Published online: 21 Nov 2011.

EUROPEAN JOURNAL OF DEVELOPMENTAL PSYCHOLOGY
2011, 8 (6), 713–729

Psychology Press
Taylor & Francis Group

An introduction to Bayesian model selection for evaluating informative hypotheses

Rens van de Schoot[1], Joris Mulder[1], Herbert Hoijtink[1], Marcel A. G. Van Aken[2], Judith Semon Dubas[2], Bram Orobio de Castro[2], Wim Meeus[3], and Jan-Willem Romeijn[4]

[1]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
[2]Department of Developmental Psychology, Utrecht University, Utrecht, The Netherlands
[3]Research Centre Adolescent Development, Utrecht University, Utrecht, The Netherlands
[4]Department of Philosophy, Groningen University, Groningen, The Netherlands

Most researchers have specific expectations concerning their research questions. These may be derived from theory, empirical evidence, or both. Yet despite these expectations, most investigators still use null hypothesis testing to evaluate their data, that is, when analysing their data they ignore the expectations they have. In the present article, Bayesian model selection is presented as a means to evaluate the expectations researchers have, that is, to evaluate so called informative hypotheses. Although the methodology to do this has been described in previous articles, these are rather technical and have mainly been published in statistical journals. The main objective of the present article is to provide a basic introduction to the evaluation of informative hypotheses using Bayesian model selection. Moreover, what is new in comparison to previous publications on this topic is that we provide guidelines on how to interpret the results. Bayesian evaluation of informative hypotheses is illustrated using an example concerning psychosocial functioning and the interplay between personality and support from family.

Correspondence should be addressed to Rens van de Schoot, Utrecht University, PO Box 80.140, NL-3508 TC, Utrecht, The Netherlands. E-mail: a.g.j.vandeschoot@uu.nl

Supported by a grant from the Netherlands organization for scientific research: NWO-VICI-453-05-002.

http://www.psypress.com/edp http://dx.doi.org/10.1080/17405629.2011.621799

Keywords: Bayesian model selection; Inequality constraints; Informative hypothesis; Bayes factors; Psychosocial functioning; Personality.

Statistical hypothesis evaluation has moved beyond simply testing the traditional null hypothesis: "nothing is going on" (Van de Schoot, Hoijtink, & Romeijn, 2011). Recent developments in statistics have rendered tools that enable the direct evaluation of predetermined informative hypotheses. In this paper we will introduce one such development: the evaluation of informative hypotheses formulated with inequality constraints using Bayesian model selection. Previous literature about the use of Bayesian statistics to evaluate hypotheses is available (Berger & Pericchi, 1996; Edwards, Lindman, & Savage, 1963; Gallistel, 2009; Lee, 2007; Myung & Pitt, 1997; O'Hagan, 1995; Perez & Berger, 2002; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), and also about a Bayesian evaluation of informative hypotheses (Hoijtink, 1998, 2000, 2001; Hoijtink, Klugkist, & Boelen, 2008; Klugkist & Hoijtink, 2007; Klugkist, Laudy, & Hoijtink, 2005; Kuiper, Klugkist, & Hoijtink, 2010; Laudy, Boom, & Hoijtink, 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink, & Klugkist, 2010; Mulder, Klugkist et al., 2009; Van de Schoot, 2010). There are also applied articles emerging in the field of the social sciences where a Bayesian evaluation of informative hypotheses has been used (Kammers, Mulder, De Vignemont, & Dijkerman, 2009; Laudy et al., 2005; Meeus, Van de Schoot, Keijsers, Schwartz, & Branje, 2010; Meeus, Van de Schoot, Klimstra, & Branje, 2011; Van de Schoot & Wong, 2011; Van Well, Kolk, & Klugkist, 2008). However, an easy-to-read introduction to the evaluation of informative hypotheses using Bayesian model selection and general guidelines on how to interpret the results are still lacking and this is exactly what we will provide in the current paper. We first introduce informative hypotheses and Bayesian model selection. Then we provide guidelines on how to use the methodology and how to interpret the results. Finally, we provide a real-life example where we demonstrate the guidelines.

WHAT ARE INFORMATIVE HYPOTHESES?

Informative hypotheses contain information about the ordering of means, regression coefficients or any other statistical parameter and can be constructed using the following constraints: (1) larger than, denoted by "$>$"; (2) smaller than, denoted by "$<$"; and (3) equal to, denoted by "$=$". Such expectations about the ordering of parameters can stem from previous studies, a literature review or even academic debate. If no information is available about the ordering of two parameters, they are separated by ",". An informative hypothesis can consist of combinations of constraints

among, for example, a set of means (denoted by μ). An example is the hypothesis H_1: $\{\mu_1, \mu_2\} < \mu_3 = \mu_4$, where groups 1 and 2 are both expected to have smaller mean scores than groups 3 and 4. Also, groups 1 and 2 are not, but groups 3 and 4 are restricted to have the same value. (In)equality constraints can also be used between (combinations of) means and a threshold to explicit effect sizes, for example, H_2: $\mu_1 - \mu_2 > .20$; $\mu_3 - \mu_4 < .50$, where the difference between the means of groups 1 and 2 is expected to be larger than .20 and where the difference between groups 3 and 4 is expected to be smaller then .50. If no constraints are imposed on any of the means, and any ordering is equally likely, the unconstrained hypothesis H_3: $\mu_1, \mu_2, \mu_3, \mu_4$ is obtained.

BAYESIAN STATISTICS

In the current paper we show how to analyse informative hypotheses about a set of means using the (free) software as described in Mulder, Klugkist et al. (2009; see also Mulder, Hoijtink et al., 2010). This software can deal with (M)AN(C)OVA, regression analysis, repeated-measure analyses with time-varying and time-invariant covariates. Other (free) software is available for ANCOVA (Klugkist et al., 2005); latent class analyses (Hoijtink, 1998, 2001; Laudy et al., 2005) and order restricted contingency tables (Laudy & Hoijtink, 2007). A first attempt can best be made using the software programme "confirmatory ANOVA" (Kuiper & Hoijtink, 2010; Kuiper et al., 2010). Readers interested in the software can visit www.tinyurl.com/informativehypotheses.

Example

Consider a very simple example to set the stage for introducing the methodology. Suppose the research question is whether the mean score on a dependent variable, say externalizing behavioural problems, differs between two groups, say over- (denoted by μ_O) and under-(denoted by μ_U) controlled adolescents. Furthermore, suppose the first hypothesis (H_A) postulates that there is no restriction between the means (that is, any combination of means is admissible). The second hypothesis (H_B) postulates that the difference between both groups is smaller than 0.10 times the variance to reflect a small effect size. The third hypothesis (H_C) postulates that over-controlled adolescents score lower on externalizing problem behaviour than under-controlled adolescents. Formally, the three hypotheses in this simple example are:

$$\begin{aligned} &H_A: \mu_O, \mu_U; \\ &H_B: \mu_O - \mu_U < .10 \times \sigma; \\ &H_C: \mu_O < \mu_U. \end{aligned} \tag{1}$$

Of course, these hypotheses can be evaluated using classical null hypothesis testing, or one-sided hypothesis testing. However, when there are more groups, more variables, or more constraints, null hypothesis testing is not the appropriate tool see Van de Schoot, Hoijtink et al. (2011) for a detailed discussion. Moreover, *p*-values are fundamentally incompatible with measures of evidence.

Bayesian model selection consist of four components.

Admissible parameter space

The first component is the "admissible parameter space", which results from the (in)equality constraints imposed on the means (see, e.g., Halpern, 2003, p. 12, for a philosophical introduction to parameter space). Let the squares in Figure 1 represent the total parameter space for all possible combinations of μ_O and μ_U in the population.

To keep our explanation simple, we assume the parameter space is bounded between 1–3, but in fact this is not the case, see the appendix for technical details.

Now, let the *admissible* parameter space be the total of all possible combinations of μ_O and μ_U that satisfy the restrictions of each of the hypotheses (i.e., H_A, H_B, H_C). For H_A, every combination of μ_O and μ_U is permitted and, therefore, the admissible parameter space of H_A is equal to the total parameter space (left-hand panel of Figure 1). For H_B, μ_O and μ_U only a small band is allowed where $\mu_O - \mu_U < .10 \times \sigma$. For H_C, only combinations of μ_O and μ_U are permitted in which μ_O is smaller than μ_U, which results in the lower triangle in the right-hand panel of Figure 1. Note that, with respect to the admissible parameter space, the hypotheses can be ordered from a small parameter space to a large parameter space: H_B, H_C, H_A.

Within the admissible parameter space a prior distribution needs to be specified, which is a key characteristic of Bayesian analyses. The

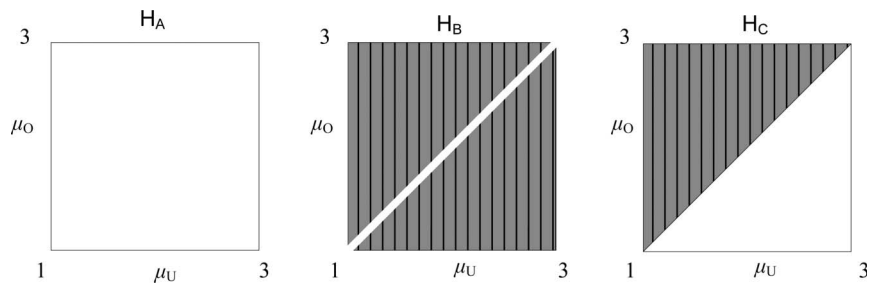

Figure 1. Admissible parameter space.

methodology employs a so called encompassing prior approach; see the appendix. The actual specification of this prior distribution is not considered to be the topic of this paper and the interested reader is referred to Mulder, Hoijtink et al. (2010) and Mulder, Klugkist et al. (2009). Note that the prior distribution is set to default in the software.

Likelihood

The second component is the likelihood of specific values of the parameters, which is the representation of the information about the means in the data set (see, e.g., Lynch, 2007, pp. 36–37). In Figure 2 an illustrative likelihood function is plotted as a function of μ_O and μ_U. The higher this surface, the more likely the corresponding combination of μ_O and μ_U in the population becomes. In this hypothetical example the sample means are 1.50 ($SD = 0.33$) for μ_O and 1.61 ($SD = 0.39$) for μ_U. So, given the data, the combination $\mu_O = 1.50$ with $\mu_U = 1.61$ is the most plausible, or the most likely combination of values for the population means. As can be seen in Figure 2, the likelihood function achieves its maximum for this combination. Other combinations of means are less likely. For example, the value of the likelihood function is much lower for the combination $\mu_O = 0.50$ and $\mu_U = 2.10$ and hence this combination of values is less likely to be the population values.

Marginal likelihood

The third component is the marginal likelihood (e.g., Chib, 1995; Kass & Raftery, 1995), which is a measure for the degree of support for each

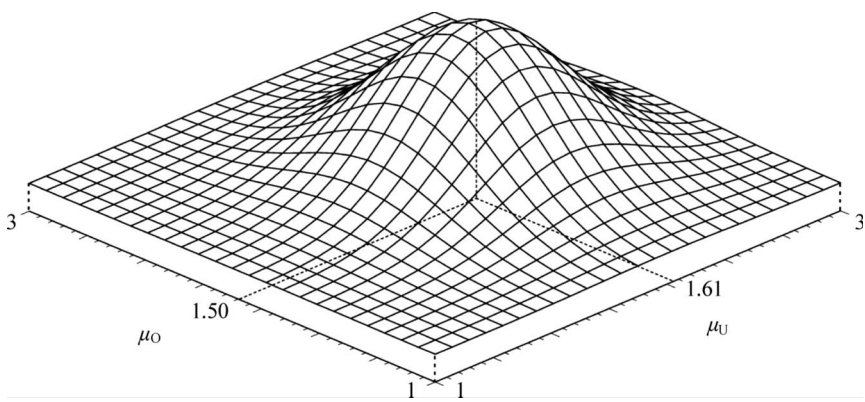

Figure 2. The likelihood function plotted as a function of μ_O and μ_U.

hypothesis provided by the data. The marginal likelihood is approximately equal to the average height of the likelihood function within the admissible parameter space. Let us elaborate on this.

Recall that Figure 1 presents the admissible parameter space for each hypothesis and Figure 2 displays the likelihood as a function of μ_O and μ_U. Both pieces of information are combined in Figure 3. The likelihood function in Figure 2 is now presented as a contour plot in Figure 3. The maximum value of the likelihood is located in the centre of the smallest circle. Remember that as you move away from this centre, the value of the likelihood of the combination of population means of μ_O and μ_U becomes smaller.

Because the admissible parameter space for H_A is equal to the total parameter space, the marginal likelihood of H_A can be computed as the average value of the likelihood in the total parameter space. This value is only meaningful in comparison to the marginal likelihood values of the other hypotheses under investigation. For H_B the average likelihood value is computed with respect to the diagonal in Figure 3 and for H_C, the average likelihood value is computed in the lower triangle in Figure 3. The marginal likelihood values are $H_A = 2.83\ e^{-67}$; $H_B = 1.81\ e^{-68}$; $H_C = 5.71\ e^{-67}$. As can be seen, H_C has the highest value, followed by H_A and then H_B.

Note that model selection can be done using the well-known model selection criteria Akaike Information Criterion (AIC) (Akaike, 1981) and Bayesian Information Criterion (BIC) (Schwartz, 1978). These selection criteria also combine fit and complexity to determine the support for a particular model. Note that the BIC is also a Bayesian selection criterion since it is derived as an approximation to the full Bayes factor. However, in contrast to Bayesian model selection both the AIC and BIC are as yet unable to deal with hypotheses specified using inequality constraints. The problem is that for the hypothesis $\mu1$, $\mu2$, $\mu3$, $\mu4$ are two distinct parameters, but is unclear how many distinct parameters there are for the hypothesis $\mu1 < \mu2 < \mu3 < \mu4$.

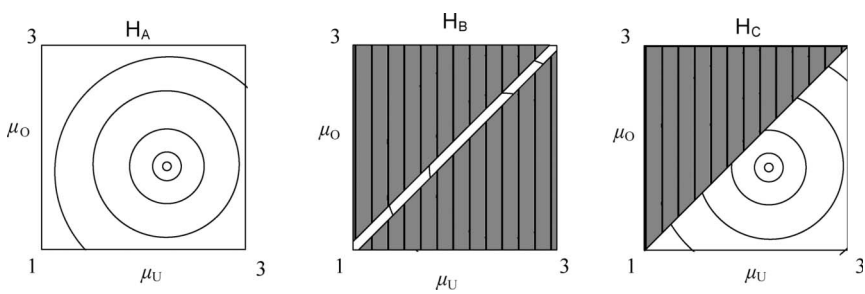

Figure 3. Likelihood function of the data within the admissible parameter space.

If we take a closer look at the plots in Figure 3, we can also observe model fit and model size, which are important components of the marginal likelihood. Many high-likelihood values are located within the admissible parameter space of H_C and H_A, but not in H_B. This indicates a good model fit for H_C and H_A, but not for H_B. Moreover, H_C has a smaller admissible parameter space compared with H_A and is therefore less complex. Furthermore, note that the likelihood values in the upper triangle in H_C are low and are not taken into account in the computation of the marginal likelihood for H_C, but are taken into account in the computation of the marginal likelihood for H_A. Consequently, the average likelihood value of H_C, and hence its marginal likelihood, is larger than the average likelihood, marginal likelihood, value of H_A. Although H_B is very parsimonious, only small likelihood values are within the admissible parameter space, which implies a poor model fit. The admissible parameter space is smallest for H_B, the marginal likelihood is smaller than that of H_A and H_C because of the ''poor'' model fit.

In sum, the marginal likelihood rewards a hypothesis with the correct (in)equality constraints. This is because the average likelihood value is higher when many small likelihood values are not taken into account. The smaller the parameter space, the less complex a model becomes. Therefore, the methodology combines model fit and model size of a hypothesis.

Bayes factors

As was shown by Klugkist et al. (2005; see also Klugkist & Hoijtink, 2007), informative hypotheses can be compared using the ratio of two marginal likelihood values, resulting in Bayes factors (denoted by BF). See Kass and Raftery (1995), for a statistical discussion of the Bayes factor and see the appendix for technical details on how model fit and model complexity are taken into account. The outcome represents the amount of evidence in favour of one hypothesis compared with another hypothesis. The results may be interpreted as follows: $BF = 1$ states that the two hypotheses are equally supported by the data; $BF > 1$ states that the support for one hypothesis is higher than for another hypothesis.

In our simple example, the BF for H_C compared to H_A can be obtained from the marginal likelihoods of both hypotheses:

$$BF_{CA} = \frac{M_C}{M_A} = \frac{5.71e^{-67}}{2.83e^{-67}} \approx 2. \tag{2}$$

For BF_{CB} the result is:

$$BF_{BA} = \frac{M_B}{M_A} = \frac{1.81e^{-68}}{5.71e^{-67}} \approx 0.031. \tag{3}$$

Now the Bayes factor between H_B and H_C can be computed by:

$$BF_{BC} = \frac{BF_{CA}}{BF_{BA}} = \frac{2}{0.031} \approx 64.51. \quad (4)$$

In conclusion, H_C receives two times more support from the data then H_A and 64 times more support than H_B. Note that the BFs can also be computed using a measure for fit and complexity directly, this is elaborated in the appendix. Recall that Bayes factors provide a direct quantification of the support in the data for the constraints imposed on the means. With support we mean: the trade-off between model size and model fit. Every researcher will agree that 31 times more support seems considerable while, for example. 1.04 times as much support does not. We refrain from providing real cut-off scores because we want to avoid creating arbitrary decision rules. Remember the famous quote about *p*-values: "... surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277). The conclusion based on Bayes factors remains subjective, just like the interpretation of other measures for support, for example, odd ratios.

GUIDELINES

When evaluating a set of informative hypotheses using Bayesian model selection, we recommend the following three-step procedure.

Step 1

In the first step the informative hypotheses have to be formulated. That is, the expected ordering of the parameters needs to be specified. If there are conflicting expectations, multiple informative hypotheses may be specified. The constraints and the observed data are the input for the software.

Step 2

After running the software, each informative hypothesis under investigation is provided with a BF against the unconstrained hypothesis. If this BF turns out to be larger than 1, it can be concluded that there is support from the data in favour of that particular informative hypothesis. If the $BF < 1$, it can be concluded that there *no* support in the data for the informative hypothesis. This procedure should be repeated for all informative hypotheses under investigation. The reason for calculating these BFs, is to enable inspection of the overall model fit of the hypotheses under

investigation. In other words, you do not want to perform model selection among poor hypotheses. Subsequently, the informative hypotheses can be divided into a set of ''supported'' hypotheses and a set of ''unsupported'' hypotheses.

Step 3

In the third step, all the informative hypotheses of interest are compared with one another (these might include ''unsupported'' hypotheses if you want). Next, all mutual BFs can be computed. However, if many informative hypotheses are considered, it is not practical to present the BFs for all possible comparisons. Instead you can provide the BFs comparing the hypothesis with the largest support of Step 1 against each of the others.

PSYCHOLOGICAL FUNCTIONING, PERSONALITY AND SUPPORT FROM FAMILY

Introduction

Van Aken and Dubas (2004) investigated whether psychosocial functioning is the result of the interplay between personality and support from family. The problem behaviour list was used to obtain parental reports on adolescents' behavioural problems. Three subscales were used, namely externalizing (E), internalizing (I) and social (S) problem behaviour. Personality types (R, O, U) were denoted using adolescents' self-reports on Big-Five personality markers (Gerris et al., 1998). Finally, the relational support inventory (Scholte, Van Lieshout, & Van Aken, 2001) was used to measure the support that children report they receive from their parents to obtain high (H) versus low (L) family support.

Based on personality type (R, O, U), high or low family support (H, L), $3 \times 2 = 6$ groups were constructed, see Table 1. Let μ denote the mean score on the dependent variable, then μ_{RHE} is the mean score for Resilient adolescents with High family support on the dependent variable Externalizing behaviour. To analyse this data we follow the three-step procedure described above.

Step 1

H_A states that under-controllers are expected to have the most externalizing problems and over-controllers are expected to have the most internalizing problems. Over-controllers and under-controllers are believed to score higher on social problems compared with resilient adolescents. Moreover,

TABLE 1
Groups of adolescents based on personality type, problem behaviour and support

		Problem behaviour		
		Internalizing	*Externalizing*	*Social*
Resilient	High family support	μ_{RHI}	μ_{RHE}	μ_{RHS}
	Low family support	μ_{RLI}	μ_{RLE}	μ_{RLS}
Over	High family support	μ_{OHI}	μ_{OHE}	μ_{OHS}
	Low family support	μ_{OLI}	μ_{OLE}	μ_{OLS}
Under	High family support	μ_{UHI}	μ_{UHE}	μ_{UHS}
	Low family support	μ_{ULI}	μ_{ULE}	μ_{ULS}

no constraints are specified with respect to high/low family support. The informative hypothesis H_A can be formulated as:

$$\begin{aligned} &(\mu_{RHE}, \mu_{RLE}, \mu_{OHE}, \mu_{OLE}) < (\mu_{UHE}, \mu_{ULE}) \\ &H_A\text{: } (\mu_{RHI}, \mu_{RLI}, \mu_{UHI}, \mu_{ULI}) < (\mu_{OHI}, \mu_{OLI}) \\ &(\mu_{RHS}, \mu_{RLS}) < (\mu_{OHS}, \mu_{OLS}, \mu_{UHS}, \mu_{ULS}). \end{aligned} \tag{5}$$

H_B states, additionally to H_A, that resilient adolescents function best in all psychosocial domains in comparison with the other two types of adolescents. Hence, the informative hypothesis H_B contains two additional constraints in comparison to H_A:

$$\begin{aligned} &(\mu_{RHE}, \mu_{RLE}) < (\mu_{OHE}, \mu_{OLE}) < (\mu_{UHE}, \mu_{ULE}) \\ &H_B\text{: } (\mu_{RHI}, \mu_{RLI}) < (\mu_{UHI}, \mu_{ULI}) < (\mu_{OHI}, \mu_{OLI}) \\ &(\mu_{RHS}, \mu_{RLS}) < (\mu_{OHS}, \mu_{OLS}, \mu_{UHS}, \mu_{ULS}). \end{aligned} \tag{6}$$

Previous research also indicates that it is the combination of personality type and the quality of social relationships that determines the risk level for experiencing more problem behaviour. Therefore, additional constraints are constructed for the third expectation (H_C). Over- and under-controllers with high perceived support from parents are expected to function better in psychosocial domains than those with low perceived support. For the resilient group, the level of support from parents is not related to problem behaviour. The constraints for informative hypothesis H_C are:

$$\begin{aligned} &(\mu_{RHE} = \mu_{RLE})(\mu_{OHE}, \mu_{OLE}) < (\mu_{UHE} < \mu_{ULE}) \\ &H_C\text{: } (\mu_{RHI} = \mu_{RLI}) < (\mu_{UHI} < \mu_{ULI}) < (\mu_{OHI} < \mu_{OLI}) \\ &(\mu_{RHS}{=}\mu_{RLS}) < (\mu_{OHS} < \mu_{OLS})(\mu_{UHS} < \mu_{ULS}). \end{aligned} \tag{7}$$

Step 2

The second step involves comparing H_A, H_B, and H_C with the unconstrained hypothesis, H_U. The results, see the second column of Table 2, show that all informative hypotheses have a BF > 1. For example, the BF between H_A and H_U is 30.28, indicating that H_A receives 30.28 times more support than H_U. From these BFs, it can be concluded that each of the hypotheses H_A, H_B, and H_C have a good model fit.

Step 3

As we showed before, mutual BFs can be computed using the results of Step 1. The BF of H_B against H_A is given by $BF_{BA} = BF_{BU}/BF_{AU} = 64.20/30.28 = 2.12$ as explained in Equations 2–4. The support for H_B is about twice as strong as for H_A. From the analyses it can be concluded that there is evidence in favour of H_A, but we think 2.12 times as much support is not much. So, our subjective conclusion might be that additional to the constraints of H_A, there is some evidence that resilient adolescents score lower on externalizing behaviour than over-controlled adolescents and that resilient adolescents score lower on internalizing behaviour than under-controlled adolescents.

The BF of H_C versus H_B, see the fourth column in Table 2, shows that there is much support in favour of H_C compared with either H_A or H_B. For example, the BF for H_C against H_B is 21.79; in other words there is approximately 21 times as much support for H_C as for H_B. From this analysis it can be concluded that the additional constraints of H_C shown in Equation 4 are a meaningful addition to the constraints of H_B.

The results of Bayesian model selection for the example relating to personality types and problem behaviour provides strong support for the idea that it is the combination of personality type and the quality of social relationships that puts adolescents at risk of greater problem behaviour.

As was correctly noticed by one of the reviewers, it can be illustrative to provide more information than just the Bayes factors. Figure 4 displays the

TABLE 2
Results of Bayesian model selection for the example of Van Aken and Dubas (2004)

Expectation	*BF**	*BF***	*BF****
H_A	30.28	1	–
H_B	64.20	2.12	1
H_C	1399.00	–	21.79

Notes: *BF compared with the unconstrained hypothesis; **BF between H_A and H_B; ***BF between H_B and H_C.

TABLE 3
Means and standard deviation (in parentheses) for the example of Van Aken and Dubas (2004)

		Problem behaviour		
		Internalizing	*Externalizing*	*Social*
Resilient	High family support (n = 137)	1.88 (0.43)	1.49 (0.32)	1.69 (0.43)
	Low family support (n = 73)	1.93 (0.47)	1.63 (0.38)	1.78 (0.48)
Over	High family support (n = 76)	2.04 (0.43)	1.43 (0.24)	1.77 (0.48)
	Low family support (n = 82)	2.17 (0.47)	1.57 (0.38)	1.93 (0.49)
Under	High family support (n = 72)	2.07 (0.57)	1.52 (0.36)	1.84 (0.52)
	Low family support (n = 135)	2.13 (0.53)	1.66 (0.39)	1.94 (0.56)

posterior distributions of the maximum likelihood estimates of the means as shown in Table 3, the posterior means and their credibility intervals. The interpretation of a Bayesian 95% credibility interval is that, for example, there is a 0.95 probability that the mean score for Resilient adolescents with High family support on the dependent variable Externalizing behaviour lies in the interval from 1.84 to 1.97 (see the plot in the upper left corner of Figure 4). These intervals are often used in practice to decide whether means differ from zero or from other means. It can, for example, be seen that the posterior mean of μ_{RHE} is 1.90 and the posterior mean of μ_{RHS} is 1.71. The credibility intervals do not overlap and, consequently, the hypothesis $\mu_{RHE} = \mu_{RHS}$ can be rejected.

DISCUSSION

In the current paper we have shown that Bayesian model selection is a useful tool when evaluating informative hypotheses. The resulting Bayes factor quantifies the amount of support received from the data for each informative hypothesis. We have offered an introduction to the methodology for non-statisticians and we are the first to present a step-by-step approach to analysing informative hypotheses with Bayesian model selection.

The major advantage of evaluating a set of informative hypotheses is that prior information can be incorporated into an analysis. As argued by Howard, Maxwell, and Fleming (2000), replication is an indispensable tool in the social sciences. Evaluating informative hypotheses fits within this framework because results from different research papers can be translated into different informative hypotheses. The method of Bayesian model selection can provide each informative hypothesis with the degree of support supplied by the data. As a result, the plausibility of previous findings can be evaluated in relation to new data, which makes the method described in this paper an interesting tool for replication of research results.

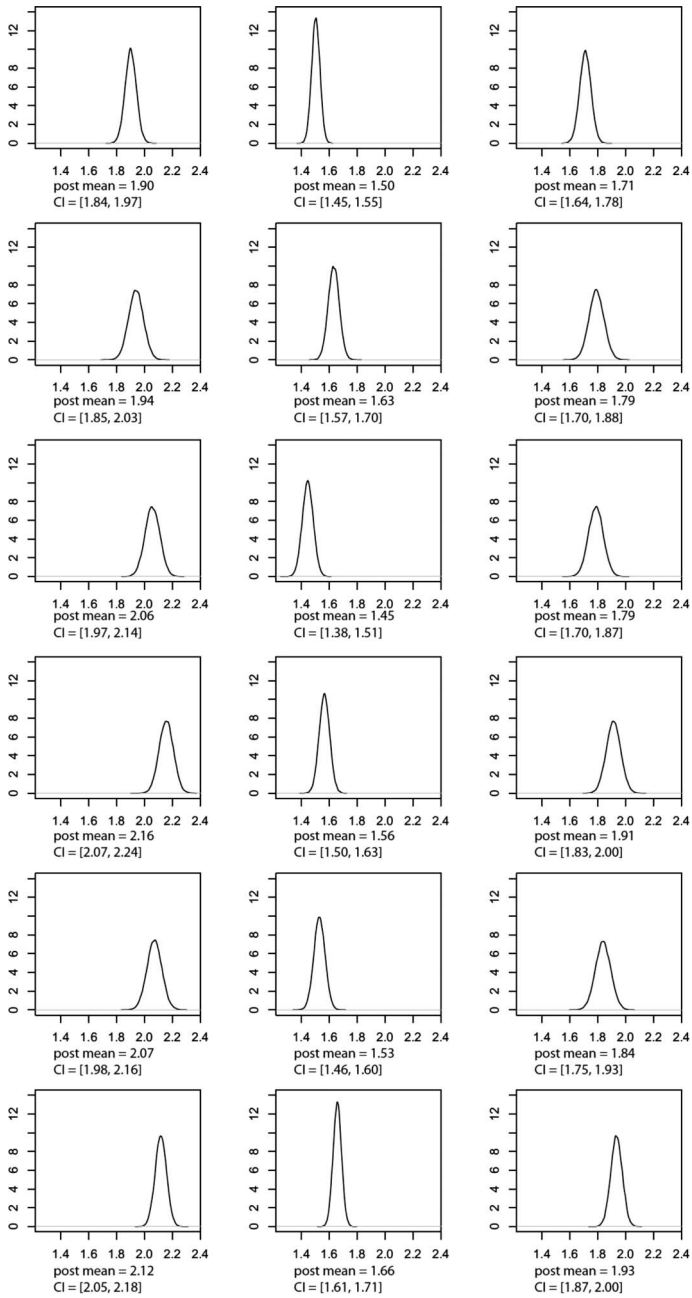

Figure 4. Posterior distributions for all groups of Table 1. *Note*: "CI" denotes the Bayesian credibility interval.

For a more detailed comparison of traditional null hypotheses testing and Bayesian evaluation of informative hypotheses we refer to Van de Schoot, Hoijtink, Mulder et al. (2011, see also Hoijtink & Klugkist, 2007; Kuiper & Hoijtink, 2010), who compared Bayesian model selection for the evaluation of inequality constrained hypotheses with classical null hypothesis testing, one-sided hypothesis testing and the best of what null hypothesis testing offers: planned comparisons. Focused analysis of variance (ANOVA) contrasts (Rosenthal, Rosnow, & Rubin, 2000) or a parametric bootstrap procedure (Van de Schoot, Hoijtink, & Deković, 2010; Van de Schoot & Strohmeier, 2011) could also be used to develop inferences regarding patterns of means. However, as soon as multiple informative hypotheses are considered, these methods are not sufficient. That is one single informative hypothesis is compared to either the classical null hypothesis or the unconstrained hypothesis.

In Bayesian statistics it would also be possible to take into consideration a priori differences in model odds as well, according to the adage "extraordinary claims require extraordinary evidence" (e.g., Jaynes, 2003). Future research should focus on how to incorporate such statements in the evaluation of informative hypotheses.

In conclusion, if researchers in psychology want to learn as much as possible from their data and if they want to judge the plausibility of expectations, Bayesian model selection, as described in the current paper, is a promising and exciting tool.

Manuscript received 11 January 2011
Revised manuscript accepted 20 June 2011
First published online 18 November 2011

REFERENCES

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3–14.

Berger, J. O., & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

Berger, J., & Pericchi, L. R. (2004). Training samples in objective Bayesian model selection. *The Annals of Statistics*, *32*, 841–869.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

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

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.

Gerris, J. R. M., Houtmans, M. J. M., Kwaaitaal-Roosen, E. M. G., Schipper, J. C., Vermulst, A. A., & Janssens, J. M. A. M. (1998). *Parents, adolescents, and young adults in Dutch families: A longitudinal study*. Nijmegen, The Netherlands: Institute of Family Studies, University of Nijmegen.

Halpern, J. Y. (2003). *Reasoning about uncertainty*. London, UK: MIT Press.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive *p*-values: Applications to educational testing. *Statistica Sinica*, *8*, 691–712.

Hoijtink, H. (2000). Posterior inference in the random intercept model based on samples obtained with Markov chain Monte Carlo methods. *Computational Statistics*, *3*, 315–336.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, *36*, 563–588.

Hoijtink, H., & Klugkist, I. (2007). Comparison of hypothesis testing and Bayesian model selection. *Quality and Quantity*, *41*, 73–91.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer.

Howard, G. S., Maxwell, S. E., & Fleming, K. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315–332.

Jaynes, E. T. (2003). *Probability theory, the logic of science*. Cambridge, UK: Cambridge University Press.

Kammers, M. P. M., Mulder, J., De Vignemont, F., & Dijkerman, H. C. (2009). The weight of representing the body: Addressing the potentially indefinite number of body representations in healthy individuals. *Experimental Brain Research*, *3*, 333–342.

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

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*, 6367–6379.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using confirmatory and exploratory approaches. *Psychological Methods*, *15*, 69–86.

Kuiper, R. M., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software*, *34*, 1–31.

Laudy, O., Boom, J., & Hoijtink, H. (2005). Bayesian computational methods for inequality constrained latent class analysis. In A. Van der Ark & M. A. C. K. Sijtsma (Eds.), *New development in categorical data analysis for the social and behavioural sciences* (pp. 63–82). London, UK: Lawrence Erlbaum Associates, Ltd.

Laudy, O., & Hoijtink, H. (2007). Bayesian methods for the analysis of inequality constrained contingency tables. *Statistical Methods in Medical Research*, *16*, 123–138.

Laudy, O., Zoccolillo, M., Baillargeon, R., Boom, J., Tremblay, R., & Hoijtink, H. (2005). Applications of confirmatory latent class analysis in developmental psychology. *European Journal of Developmental Psychology*, *2*, 1–15.

Lee, S.-Y. (2007). *Structural equation modeling: A Bayesian approach*. Chichester, UK: Wiley.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer.

Meeus, W., Van de Schoot, R., Keijsers, L., Schwartz, S. J., & Branje, S. (2010). On the progression and stability of adolescent identity formation. A five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Child Development*, *81*, 1565–1581.

Meeus, W., Van de Schoot, R., Klimstra, T., & Branje, S. (2011). Change and stability of personality types in adolescence: A five-wave longitudinal study in early-to-middle and middle-to-late adolescence. *Developmental Psychology*, *47*(4), 1181–1195.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887–906.

Mulder, J., Klugkist, I., Van de Schoot, R., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Informative hypotheses for repeated measurements: A Bayesian approach. *Journal of Mathematical Psychology*, *53*, 530–546.

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

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, *57*, 99–138.

Perez, J. M., & Berger, J. (2002). Expected posterior prior distributions for model selection. *Biometrika*, *89*, 491–511.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge, UK: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*, 1276–1284.

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.

Scholte, R. H. J., Van Lieshout, C. F. M., & Van Aken, M. A. G. (2001). Relational support in adolescence: Factors, types, and adjustment. *Journal of Research in Adolescence*, *11*, 71–94.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Van Aken, M. A. G., & Dubas, J. D. (2004). Personality type, social relationships, and problem behaviour in adolescence. *European Journal of Developmental Psychology*, *1*, 331–348.

Van de Schoot, R. (2010). *Informative hypotheses. How to move beyond classical null hypothesis testing*. Utrecht, The Netherlands: Utrecht University, PhD thesis. (Accessible at: http://igitur-archive.library.uu.nl/dissertations/2010-0909-200248/UUindex.html).

Van de Schoot, R., Hoijtink, H., & Deković, M. (2010). Testing inequality constrained hypotheses in SEM models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*, 443–463.

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W., et al. (2011). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology*, *47*, 203–212.

Van de Schoot, R., Hoijtink, H., & Romeijn, J.-W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Quantitative Psychology and Measurement*, *2*(24). doi:10.3389/fpsyg.2011.00024

Van de Schoot, R., & Strohmeier, D. (2011). Testing informative hypotheses in SEM increases power: An illustration contrasting classical hypothesis testing with a parametric bootstrap approach. *International Journal of Behavioural Development*, *35*, 180–190.

Van de Schoot, R., & Wong, T. (2011). Do antisocial young adults have a high or a low level of self-concept? *Self and Identity*. doi:10.1080/15298868.2010.517713

Van Well, S., Kolk, A. M., & Klugkist, I. (2008). The relationship between sex, gender role identification, and the gender relevance of a stressor on physiological and subjective stress responses: Sex and gender (mis)match effects. *Behavior Modification*, *32*, 427–449.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*, 158–189.

APPENDIX

Technical details of the Bayesian methodology for ANOVA

In the appendix we provide more technical details for the methodology of evaluating informative hypothesis for ANOVA models. Hypotheses with inequality constraints are nested in the unconstrained model. The Bayes

factor of an inequality constrained hypothesis, H_i, versus an unconstrained hypothesis, H_u, can be written as:

$$BF_{iu} = \frac{M_i}{M_u} = \frac{f_i}{c_i} \tag{A1}$$

where M is the marginal likelihood value, see Equation 2 in the main text. Moreover, f_i can be interpreted as the fit of the model under investigation and c_i as the complexity of H_i. Note that this representation of the BF differs from the BF in the main text. This is due to the fact that the marginal likelihood is often difficult to compute and Klugkist et al. (2005) introduced the BF in Equation A1 as an alternative.

Complexity, c_i, is measured as the proportion of the prior distribution of H_i in agreement with the constraints imposed on the parameters of interest. According to the encompassing prior approach as proposed by Klugkist et al. (2005), the prior $h(.)$ for H_u in the context of ANOVA models (with J means) is given by:

$$h(\mu_1 \ldots \mu_J, \sigma^2 | H_U) = N(\mu_1 | \mu_0, \tau_0^2) \times \cdots \times N(\mu_J | \mu_0, \tau_0^2) \times Inv - Gamma(\sigma^2 | a, b) \tag{A2}$$

where μ_0 denotes the prior mean, τ_0^2 the precision (which are the same for each mean) and a and b are respectively, the shape and the scale parameter of the inverse gamma distribution. From Equation A2, it can be derived that:

$$c_i = \int_{Hi} h(\mu_1 \ldots \mu_J, \sigma^2 | H_U) d\mu_1 \ldots \mu_J \tag{A3}$$

The posterior distribution, $g(\cdot)$, is proportional to the product of the prior and the likelihood function of the data: $g(\cdot) \propto h(\cdot) \times f(\cdot)$. If Y is observed data, model fit f_i is then given by:

$$f_i = \int_{Hi} g(\mu_1 \ldots \mu_J, \sigma^2 | Y, H_u) d(\mu_1 \ldots \mu_J) \tag{A4}$$

Stated otherwise, f_c is the proportion of the posterior distribution of H_U in agreement with H_i. For ANOVA models, the likelihood is given by:

$$f(y | \mu_1 \ldots \mu_J, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left\{ \frac{(y_i - \Sigma_{j=1}^{J} \mu_j d_{ij})^2}{2\sigma^2} \right\} \tag{A5}$$

where d_{ij} denotes whether a case belongs to group j or not.

The prior parameters, μ_0, τ_0^2, a and b of Equation A2 are obtained using training data, as is presented in Berger and Pericchi (1996, 2004), O'Hagan (1995), Perez and Berger (2002) and Mulder and colleagues (2010).