Developmental Psychology
2011, Vol. 47, No. 1, 203–212

0012-1649/11/$12.00 DOI: 10.1037/a0020957

Evaluating Expectations About Negative Emotional States of Aggressive Boys Using Bayesian Model Selection

Rens van de Schoot, Herbert Hoijtink, Joris Mulder, Marcel A. G. Van Aken, Bram Orobio de Castro, and Wim Meeus
Utrecht University, The Netherlands

Jan-Willem Romeijn
Groningen University, The Netherlands

Researchers often have expectations about the research outcomes in regard to inequality constraints between, e.g., group means. Consider the example of researchers who investigated the effects of inducing a negative emotional state in aggressive boys. It was expected that highly aggressive boys would, on average, score higher on aggressive responses toward other peers than moderately aggressive boys, who would in turn score higher than nonaggressive boys. In most cases, null hypothesis testing is used to evaluate such hypotheses. We show, however, that hypotheses formulated using inequality constraints between the group means are generally not evaluated properly. The wrong hypotheses are tested, i.e.. the null hypothesis that group means are equal. In this article, we propose an innovative solution to these above-mentioned issues using Bayesian model selection, which we illustrate using a case study.

Keywords: Bayesian model selection, informative hypothesis, power, planned comparison, one-sided hypothesis testing, aggression, emotional state

Many psychology researchers rely on regression analysis, analysis of variance, or repeated-measures analysis to answer their research questions. The default approach in these procedures is to test the classical null hypothesis that "nothing is going on": regression coefficients are zero, there are no group differences, and so on. We argue that many researchers have some very strong prior beliefs about various components outcomes of their analyses and are not particularly interested in testing a traditional null hypothesis (see also Cohen, 1990, 1994; Wagenmakers, 2007). For example, a researcher might expect that highly aggressive boys would, on average, score higher on aggressive responses towards other peers than moderately aggressive boys, who in turn would score higher than nonaggressive boys. Note that we refer to such explicit expectations as informative hypotheses.

This aforementioned explicit expectation is clearly not the same as the traditional null hypothesis: All scores for the boys are equal. Often researchers are not particularly interested in this null hypothesis. However, the average researcher specifies the traditional null hypothesis in a robotic way. Note that this is a critique of the researcher and not of the method, since classical null hypothesis testing is very useful for testing the null hypotheses if one is interested in it. Even so, there are already researchers who actually use prior beliefs directly in their data analyses (see, e.g., Kammers, Mulder, De Vignemont, & Dijkerman, 2009; Meeus, Van de Schoot, Keijsers, Schwartz, & Branje, in press; Meeus, Van de Schoot, Klimstra, & Branje, in press; Van de Schoot, Hoijtink, & Doosje, 2009; Van de Schoot & Wong, in press; van Well, Kolk, & Klugkist, 2009).

In this article, we show how subjective beliefs influence analyses in hidden ways and how they might be incorporated explicitly in data analysis. That is, we describe, by means of a case study, what can happen if a researcher has informative hypotheses and uses traditional frequentist analysis or thoughtful frequentist analysis. Subsequently, we elaborate on an alternative strategy: the evaluation of informative hypotheses by means of Bayesian model selection (Hoijtink, 1998, 2001; Hoijtink, Klugkist, & Boelen, 2008; Klugkist, Laudy, & Hoijtink, 2005; Klugkist, Laudy, & Hoijtink, 2010; Kuiper & Hoijtink, 2010; Laudy, Boom, & Hoijtink, 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink, & Klugkist, 2009; Mulder, Klugkist, Van de Schoot, Meeus, Selfhout, & Hoijtink, 2009). Furthermore, we use one of our own studies (Orobio de Castro, Slot, Bosch, Koops, & Veerman, 2003) in the area of experimental psychology to illustrate that our aim is not to disregard any specific study but to discuss a problem very common to psychological research, a problem encountered in our own research as well.

Example: Emotional State in Aggressive Boys

Orobio de Castro et al. (2003) investigated the effects of inducing a negative emotional state in aggressive boys (overall $M = 11$ years, $SD = 1.2$ year). The question was whether inducing negative emotions would make boys with aggressive behavior prob-

Rens van de Schoot, Herbert Hoijtink, and Joris Mulder, Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands; Marcel A. G. Van Aken and Bram Orobio de Castro, Department of Developmental Psychology, Utrecht University; Wim Meeus, Department of Child and Adolescent Studies, Utrecht University; and Jan-Willem Romeijn, Department of Philosophy, Groningen University, Groningen, The Netherlands.

This article was supported by a grant from The Netherlands Organisation for Scientific Research (NWO-VICI-453-05-002). Many thanks to Wenneke Hubeek for her support and for proofreading the manuscript. We would also like to thank Franz Mechsner for his feedback on the manuscript and the reviewers for their many good suggestions.

Correspondence concerning this article should be addressed to Rens van de Schoot, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands. E-mail: a.g.j.vandeschoot@uu.nl

lems attribute more aggressive responses and hostile intentions to their peers than did a group of nonaggressive boys. The authors examined three levels of aggression: high, moderate, and none.

The highly aggressive group consisted of boys referred to special education for aggressive behavior problems. Informed consent was obtained from all participants and their parents. The moderately aggressive group consisted of boys in regular education with teacher-rated externalizing behavior problem scores on the Teacher's Report Form (Achenbach, 1991; for the Dutch version, see Verhulst, Van der Ende, & Koot, 1997) in the borderline or clinical range. No socioeconomic status information was made available from the original article.

The authors induced mild negative emotions by manipulating participants' performance in a computer game. Each participant completed two conditions: a neutral-emotion condition prior to playing a computer game (*neutral*) and a negative-emotion condition following emotional manipulation after unjustly losing the game (*negative*). The authors assessed hostile intent attributions and aggressive responses to other peers by presenting the boys with eight vignettes concerning ambiguous provocation by peers, for example:

> Imagine: You and a boy in your class are taking turns at a computer game. Now it's your turn, and you are doing great. You are reaching the highest level, but you only have one life left. You never came this far before, so you are trying very hard. The boy you are playing with watches the game over your shoulder. He sees how far you have come. Then he shouts "Watch out! You've got to be fast now!" and he pushes a button. But it was the wrong button, and now you have lost the game!

Two open-ended questions were asked directly after listening to each vignette: (a) why the provocateur in the vignette acted the way he did; and (b) how the participants would respond were they to actually experience the events portrayed in the vignette. Answers to the first question were coded as benign, accidental, ambiguous, or hostile. The reactions of the boys to the second question were coded as aggressive, coercive, solution attempt, or avoidant. The authors calculated respective scores for hostile intentions and responses by counting the number of vignettes in each condition with a hostile or an aggressive response to the questions.

Expectations

The first expectation (A) was that negative emotion manipulation would invoke more hostile intentions and aggressive responses at all levels of aggression. This expectation was made on the basis of Dodge's (1985) hypothesis that a negative emotional state makes children more prone to attribute hostile intentions to other children with whom they interact. The constraints corresponding to the informative hypothesis $H_{\mathrm{A,host}}$ in relation to hostile attribution are displayed in Table 1. It can be seen, for example, that the mean score for nonaggressive boys in the neutral condition is expected to be lower than the mean score for nonaggressive boys in the negative condition, $M_{neu,non} < M_{neg,non}$. Note that the same constraints hold for aggressive responses ($H_{\mathrm{A,aggr}}$).

A second expectation (B) was that emotion manipulation would influence aggressive boys more than it would less aggressive boys. Consequently, the tendency to attribute more hostile intentions to peers in ambiguous situations was expected to increase more in highly aggressive boys than in moderately aggressive and nonaggressive boys. As was argued by Orobio de Castro et al. (2003), this hypothesis seems plausible, given the fact that many children with aggressive behavior problems have histories of abuse, neglect, and rejection (Coie & Dodge, 1998). As a result, these highly aggressive boys exhibit a greater tendency to attribute hostile intentions to peers in ambiguous situations than nonaggressive boys do (see also, Orobio de Castro, Veerman, Koops, Bosch, & Monshouwer, 2002). The constraints corresponding to the informative hypothesis for hostile attribution ($H_{\mathrm{B,host}}$) are displayed in the middle of Table 1. These constraints imply, for example, that the difference between the negative and neutral conditions is smaller for the nonaggressive group than for the moderately aggressive group, $[M_{neu,non} - M_{neg,non}] < [M_{neu,mod} - M_{neg,mod}]$. The same constraints also hold for aggressive responses ($H_{\mathrm{B,aggr}}$).

A third expectation (C) was a combination of expectations A and B. The authors expected that negative emotion manipulation would invoke more hostile intentions and aggressive responses at all levels of aggression and, at the same time, that emotion manipulation would influence aggressive boys more than less aggressive boys (Orobio de Castro et al., 2003). The difference between the neutral and the negative condition would be larger if boys were

Table 1
Constraints for Hypotheses A, B, and C for Hostile Attribution

		Aggression level				
Hypothesis	Condition	No aggression		Moderate		High
$H_{\mathrm{A,host}}$	Neutral	$M_{neu,non}$		$M_{neu,mod}$		$M_{neu,high}$
		$\wedge$		$\wedge$		$\wedge$
	Negative	$M_{neg,non}$		$M_{neg,mod}$		$M_{neg,high}$
$H_{\mathrm{B,host}}$		$M_{neg,non} - M_{neu,non}$	$<$	$M_{neg,mod} - M_{neu,mod}$	$<$	$M_{neg,high} - M_{neu,high}$
$H_{\mathrm{C,host}}$	Neutral	$M_{neu,non}$		$M_{neu,mod}$		$M_{neu,high}$
		$\wedge$		$\wedge$		$\wedge$
	Negative	$M_{neg,non}$		$M_{neg,mod}$		$M_{neg,high}$
				and		
		$M_{neg,non} - M_{neu,non}$	$<$	$M_{neg,mod} - M_{neu,mod}$	$<$	$M_{neg,high} - M_{neu,high}$

Note. M indicates a mean score for an aggression level within a condition (e.g., $M_{neu,non}$ is the mean score for non-aggressive boys in the neutral condition).

more aggressive. The hypotheses $H_{C,host}$ and $H_{C,aggr}$ combine the constraints presented in the upper part of Table 1 with the constraints presented in the middle of Table 1.

The research question we investigate in the current article is which of these three informative hypotheses, H_A, H_B, or H_C, is best supported by the data. We try to answer this research question using traditional frequentist analysis, thoughtful frequentist analysis, and Bayesian model selection.

Traditional Frequentist Analysis

The traditional frequentist approach, which is most often used in practice, is to analyze data like ours using traditional null hypothesis testing. In our example, we used aggressive responses and hostile intentions as dependent variables in two analyses of variance (ANOVA) with level of aggression (high, moderate, and no aggression) as a between-participants factor and condition (neutral/negative) as a within-participants factor. Three null hypotheses could be tested for both hostile intentions and aggressive responses:

$H_{0,1}$: There is no difference among levels of aggression;

$H_{0,2}$: There is no difference between the condition means;

$H_{0,1\times2}$: There is no interaction between level of aggression and the condition.

The results of these tests are presented in Table 2 (significant results are in bold). It can be seen in this table that for both aggressive responses and hostile intentions, there appear to be significant differences between aggression level means and that there were no differences between condition means for both aggressive responses and hostile intentions. However, the only significant result for the interaction effect is found for hostile attribution (i.e., Level of Aggression × Condition). Many researchers would perform a follow-up analysis, which we also do, but we first show what happens if the informative hypotheses H_A, H_B, and H_C are evaluated using the null hypotheses $H_{0,1}$, $H_{0,2}$, and $H_{0,1\times2}$.

What Goes Wrong?

Although traditional null hypothesis testing has been the dominant research tool for the latter half of the past century, it suffers from serious complications if used in the wrong way—that is, when the null hypotheses $H_{0,1}$, $H_{0,2}$, and $H_{0,1\times2}$ are used to determine which informative hypothesis, H_A, H_B, or H_C, is best supported by the data. Let us elaborate on this.

Table 2
Results of the Two 3 × 2 Univariate Analyses of Variance

	Hostile		Aggressive	
Variable	*F*	*p*	*F*	*p*
Aggressive level (*df:* 2, 55)	2.91	**.047**	8.82	**<.001**
Condition differences (*df:* 2, 55)	1.10	.29	0.82	.36
Interaction (*df:* 2, 54)	3.18	**.049**	1.46	.24

Note. Significant results are in bold.

The first and most vital problem is that there is no straightforward relationship between the informative hypotheses under investigation and the null hypotheses that are actually being tested. Orobio de Castro et al. (2003) were not interested in testing the hypotheses $H_{0,1}$, $H_{0,2}$, and $H_{0,1\times2}$ that were tested in the ANOVA. Although Wainer (1999) argues in "One Cheer for Null Hypothesis Significance Testing" that the null hypothesis can be useful in some cases, many researchers have no particular interest in the null hypothesis (see, e.g., Cohen, 1990, 1994). So why test the null hypothesis if one is not interested in it?

Furthermore, the informative hypotheses H_A, H_B, and H_C differ from the traditional alternative hypotheses: "not $H_{0,1}$," "not $H_{0,2}$", and "not $H_{0,1\times2}$." As can be seen in Table 2, some of the null hypotheses are rejected in favor of the alternative hypothesis (significant results are in bold), but what does this tell us? For example, for hostile attribution there is a main level of aggression difference and an interaction between level of aggression and condition. Does this provide any evidence that one of the three informative hypotheses is more likely than the other? Clearly, the answer is "no," because neither the null hypotheses nor the alternative hypotheses resemble any of the informative hypotheses under investigation.

In conclusion, using traditional null hypothesis testing does not result in a direct answer to the research question at hand. This issue is usually solved by a visual inspection of the sample means. When inspecting Table 3, which shows the descriptive statistics (i.e., standardized means), it appears that there is a violation of Expectation A with regard to hostile attribution: the mean of the nonaggressive group is lower in the negative condition than it is in the neutral condition, rather than higher. Does this imply that Expectation A is not supported by the data? Or is this a random deviation? The mean differences for hostile attribution between the neutral and negative condition for non-, moderate- and high-aggressive boys, presented in the lower part of Table 3, are in agreement with the constraints of Expectation B. However, does this imply that H_B is preferred over H_A? What if there had been a small deviance of the constraints imposed on the mean differences: −.45, −.46, .45? Or what if there had been a larger deviance between the mean differences: −.45, −.55, .45? When would the difference be large enough to conclude that the informative hypothesis was preferred?

Multiple Hypothesis Testing and Power

Alongside the complication of using it the wrong way, the procedure of traditional null hypothesis testing itself suffers from a number of complications. We discuss two important issues here: an increase of Type I errors due to multiple analyses and the loss of power that results from the adjustment often used to correct for these errors.

Multiple tests are typically needed to evaluate the informative hypotheses at hand, and this can be problematic (e.g., Maxwell, 2004). In our example, six *F* tests were performed. In general, multiple testing increases the family-wise error rate, which is the probability of incorrectly rejecting at least one null hypothesis of all of those tested. For example, for two independent tests and an alpha level of .05 per test, the probability of correctly concluding that both null hypotheses are not rejected is .95 × .95 = .90, and for six tests, $.95^6$ = .74. In the latter case, the probability of

Table 3
Emotion Ratings by Aggression Level and Condition

		Hostile					Aggressive				
Hypothesis	Condition	No aggression		Moderate		High	No aggression		Moderate		High
$H_{A,host}$	Neutral	0.15		0.39		−0.27	0.52		1.02		1.12
		∧		∧		∧	∧		∧		∧
	Negative	−0.20		0.43		0.18	0.47		1.08		0.93
$H_{B,host}$		−0.45	<	0.04	<	0.45	−0.05	<	0.06	<	−0.19

incorrectly rejecting at least one null hypothesis is $1 - .74 = .26$. Note that the six tests in Table 1 are not independent, but in this situation, the overall alpha level is higher than .05 as well.

A solution to the problem of Type I error inflation is to control the overall alpha level by using, for example, the often-used Bonferroni correction. For this procedure, the overall alpha level is divided by the number of tests performed. The price for using such a correction is a severe reduction in power (see Cohen, 1992). Correcting the alpha level also requires a larger sample size to maintain sufficient power, which may not always be realistic. In our running example, ethical and clinical considerations urge us to limit to an absolute minimum the number of boys with severe behavior problems who can be asked to participate in such a taxing manipulation. These sample size restrictions are evident in many studies in our field. Moreover, the Bonferroni correction is not unproblematic; the procedure is rather conservative, meaning that the smaller the alpha level, the lower the power. Improvements on the Bonferroni procedure have been developed, including the false discovery rate (Benjamini & Hochberg, 1995) or the Holm-Bonferroni method (Holm, 1979); for an overview see Hsu (1996). However, larger sample sizes are still needed in these cases, and it remain difficult to determine how the overall alpha level should be corrected with all of these methods.

For example, when using any form of correction, should the overall alpha be corrected separately for each dependent variable? Or should the overall alpha be corrected by using the total number of tests? The answers to these questions are not clear. If we were to use the Bonferroni correction $\frac{\alpha}{3}$ for our example, then the significant results for hostile attribution disappear, and the conclusion should be that there are no group main differences and that there is no interaction between group and condition. The null hypothesis cannot be rejected, but what does this say about the informative hypotheses H_A, H_B, and H_C?

For aggressive responses, aggression level differences remain significant when using $\frac{\alpha}{3}$, implying that $(M_{non,neg} + M_{non,neu}) \neq (M_{mod,neg} + M_{mod,neu}) \neq (M_{neg,high} + M_{neu,high})$, where M is the mean score of a group within a condition. A significant result would indicate that $(0.52 + 0.47 = 0.99) \neq (1.02 + 1.08 = 1.10) \neq (1.12 + 0.93 = 2.05)$, but what can we learn from this with respect to H_A, H_B, and H_C? Clearly, the answer is "not much." Even if we pursue this significant result further using post-hoc comparisons, these comparisons do not provide information about the informative hypotheses A, B, or C.

Thoughtful Frequentist Analysis

What have we learned so far? Testing the null hypotheses $H_{0,1}$, $H_{0,2}$, and $H_{0,1\times2}$ followed by a visual inspection of the data is not the appropriate tool for evaluating the informative hypotheses H_A, H_B, and H_C. If a researcher has explicit expectations in the form of inequality constraints between means, he or she might be better off using alternative procedures. In this section, we use thoughtful frequentist analysis, in other words, planned comparisons, to evaluate H_A, H_B, and H_C.

First, three one-sided *t*-tests could be performed to evaluate H_A:

$$M_{neu,non} < M_{neg,non}\ (p_{hostile} = .22/2;\ p_{aggr} = .60/2);$$

$$M_{neu,mod} < M_{neg,mod}\ (p_{hostile} = .88/2;\ p_{aggr} = .60/2);$$

$$M_{neu,high} < M_{neg,high}\ (p_{hostile} = .02/2;\ p_{aggr} = .06/2).$$

To evaluate H_B, planned comparisons could be used; a good primer is presented in Rosenthal, Rosnow, and Rubin (2000), who introduced several types of contrasts. In our example, H_B could be evaluated using the linear contrast $-1 \times [M_{neu,non} - M_{neg,non}] + 0 \times [M_{neu,mod} - M_{neg,mod}] + 1 \times [M_{neu,high} - M_{neg,high}]$. A researcher who expects a monotonic relationship can create lambda weights that represent that hypothesis (see Rosenthal et al., 2000). For now, we will use a linear increase, and since this hypothesis is also directional, we expect an increase in the difference between conditions; the resulting p value can be divided by two. The results are a significant increase for hostile attribution ($p = .008/2$) but a nonsignificant result for aggression ($p = .32/2$). Both pieces of information (i.e., the results of the one-sided *t* tests and planned comparison) need to be combined to evaluate H_C, but it is unclear how to do so.

Although the above procedure generates better results than the naive procedure presented in the previous section, there is still one major problem related to thoughtful frequentist analysis. Recall that we wanted to evaluate H_A, H_B, and H_C. Using planned comparisons, in whatever form, results again in testing the null hypothesis. These tests are clearly not the same as evaluating H_A, H_B, and H_C. A different approach is called for, and this is what we do in the next section.

Bayesian Evaluation of Informative Hypotheses

As put forward by Walker, Gustafson, and Frimer (2007, p. 366), "the Bayesian approach offers innovative solutions to some challenging analytical problems that plague research in . . . psychology" (see also Howard, Maxwell, & Fleming, 2000; Lee &

Pope, 2006; Lee & Wagenmakers, 2005). The core idea of Bayesian inferences is that a priori beliefs are updated with observed evidence and both are combined in a so-called posterior distribution. In the social sciences, however, only a few applications of Bayesian methods can be found; one good example is presented in Walker, Gustafson, and Hennig (2001). The authors used standard statistical techniques as well as a Bayesian approach to investigate consolidation and transition models in the domain of moral reasoning. The posterior distribution of reasoning across stages of moral reasoning was used to predict subsequent development. Another example is the Schulz, Bonawitz, and Griffiths (2007) study about causal learning processes in preschoolers. Bayesian inference was used in this article to provide a rationale for updating children's beliefs in light of new evidence and was used to explore how children solve problems.

Bayes in the Social Sciences

An important contribution Bayesian methods can offer to the social sciences is the evaluation of informative hypotheses formulated with inequality constraints using Bayesian model selection. Many technical papers have been published about this method in statistical journals (Hoijtink, 1998, 2001; Hoijtink et al., 2008; Klugkist et al., 2005; Klugkist et al., 2010; Kuiper & Hoijtink, 2010; Laudy et al., 2005; Laudy & Hoijtink, 2007; Mulder et al., 2009; Mulder et al., 2009). Applied psychology/social science articles that use this method to evaluate hypotheses have been published as well.

For example, in a study by van Well et al. (2008), the authors investigated whether a possible match between sex or gender role identification on the one hand and gender relevance of a stressor on the other hand would increase physiological and subjective stress responses. A first expectation represented a sex match effect; participants were expected to be most reactive in the condition that matched their sex. In a similar way, gender match, sex mismatch, and gender mismatch effects were evaluated using Bayesian model selection software.

Another example is the study by Meeus et al. (in press). In this study, Bayesian model selection was used to evaluate the plausibility of certain patterns of increases and decreases in identity status membership on the progression and stability of adolescent identity formation. Moreover, expected differences in prevalence of identity statuses between early-to-middle and middle-to-late adolescents and males and females were evaluated. In sum, Bayesian model selection as described in, for example, Hoijtink et al. (2008), is gaining attention and is a flexible tool that can deal with several types of informative hypothesis.

The major advantage of evaluating a set of informative hypothesis using Bayesian model selection is that prior information can be incorporated into an analysis. As was argued by Howard, Maxwell, and Fleming (2000), replication is an important and indispensible 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 provided 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 article an interesting tool for replication of research results.

Another advantage of evaluating informative hypotheses is that more power is generated with the same sample size. An increase in power is achieved because using the data to directly evaluate H_A, H_B, and H_C by directly evaluating H_A versus H_B versus H_C is more straightforward than testing several null hypotheses that are not directly related to the hypotheses of interest. Besides, when H_A versus H_B versus H_C are directly evaluated, there is no need to deal with contradictory results or problems arising as a result of multiple testing.

Software

In this study, we analyzed the informative hypotheses of our example using the software presented in Mulder et al. (2009; see also Mulder et al., 2009). The method described in this article can be used to deal with many complex types of (in)equality constraints in multivariate linear models, for example, multivariate analysis of covariance (MANCOVA), regression analysis, and repeated-measure analyses with time varying and time in-varying covariates. A typical example of an informative hypothesis in the context of regression analysis can be found in Deković, Wissink, and Meijer (2004), who hypothesized that adolescent disclosure is the strongest predictor of antisocial behavior, followed by either a negative or positive relation with the parent.

Software is also available for evaluating informative hypotheses in AN(C)OVA models (Klugkist et al., 2005; Kuiper & Hoijtink, 2010) and latent class analysis (Hoijtink, 1998, 2001; Laudy et al. 2005), as well as order-restricted contingency tables (Laudy & Hoijtink, 2007; see also Klugkist et al., 2010). Readers interested in this software can visit http://www.fss.uu.nl/ms/informativehypothesis. Users of the software need only provide the data and the set of constraints; the Bayes factors are computed automatically by the software. A first attempt in analyzing data can best be made by using the software program "confirmatory ANOVA" (Kuiper, Klugkist, & Hoijtink, 2010). We refer to the book of Hoijtink et al. (2008) as a first step for interested readers.

Introduction to Bayesian Model Selection

In this section, we provide a brief introduction to the evaluation of informative hypotheses formulated with inequality constraints using Bayesian model selection. The main ideas are introduced below; we refer interested readers to Lynch (2007) for a general introduction to Bayesian analysis and to Gelman, Carlin, Stern and Rubin (2004) for a technical introduction to Bayesian analysis. For incorporating inequality constraints in the context of Bayesian model selection, we refer interested readers to Hoijtink et al. (2008).

Bayes Factor

As was shown by Klugkist et al. (2005), informative hypotheses can be compared using the ratio of two marginal likelihood values, which is a measure for the degree of support for each hypothesis provided by the data (see, e.g., Hoijtink et al., 2008). This ratio results in the Bayes factor; see Kass and Raftery (1995) for a statistical discussion of the Bayes factor. The outcome represents the amount of evidence in favor of one hypothesis compared with another hypothesis.

Returning to our example of Orobio de Castro et al. (2003), the informative hypotheses H_A, H_B, and H_C can be evaluated using Bayesian model selection. To do so, we first compare these informative hypotheses to a so-called unconstrained hypothesis, denoted by H_{unc}. A hypothesis is unconstrained if no constraints are imposed on the means. The comparison with H_{unc} is made because it is possible that all informative hypotheses under investigation do not provide an adequate description of the population from which the data were sampled. In that case, the unconstrained hypothesis will be favored by Bayesian model selection. Hence, Bayesian model selection protects a researcher against incorrectly choosing such a "bad" hypothesis.

As was shown by Klugkist et al. (2005), the Bayes factor (BF) of $H_A$versus H_{unc} can be written as

$$BF_{A,unc} = \frac{f_i}{c_i}, \quad (1)$$

where f_i can be interpreted as a measure for model fit and c_i as a measure for model complexity of H_A. The Bayes factor of H_A versus H_B can be written as:

$$BF_{A,B} = \frac{BF_{A,unc}}{BF_{B,unc}}. \quad (2)$$

The Bayes factor in Equation 2 combines model fit and complexity and represents the amount of evidence, or support from the data, in favor of one hypothesis (say, H_A) compared to another hypothesis (say, H_B).

The results may be interpreted as follows: $BF_{A,B} = 1$ states that the two hypotheses are equally supported by the data; $BF_{A,B} = 10$ states that the support for H_A is 10 times stronger than the support for H_B; $BF_{A,B} = 0.25$ states that the support for H_B is four times stronger than the support for H_A. Note that there is no cut-off value provided; we return to this issue in the next section, but let us first reanalyze our example elaborate on f_i and c_i.

Example Reconsidered

To reanalyze the data of Orobio de Castro et al. (2003), we computed the Bayes factors using two analysis of variance models, one for hostile attribution and one for aggressive response. The results are presented in Table 4.

For hostile attribution, the $BF_{A,unc}$ of $H_A$compared with H_{unc} is 0.24. This implies that H_A is not better than the unconstrained hypothesis and is consequently not supported by the data (accounting for model fit and complexity). The $BF_{B,unc}$ of H_B compared with H_{unc} is 4, indicating that support from the data is 4 times stronger for H_B than for H_{unc}. The $BF_{C,unc}$ indicates that support from the data is 1.5 times stronger for $H_C$than for H_{unc}. In sum, only H_B and H_C are supported by the data.

Using these results, one can compute a Bayes factor between two informative hypotheses. The resulting Bayes factor is equal to the ratio of the Bayes factor for each informative hypothesis compared with the unconstrained hypothesis using Equation 2. The $BF_{B,C}$ for hostile attribution between H_B and H_C is $\frac{4}{1.5} = 2.66$, which means that the support for H_B is 2.66 times stronger than the support for H_C. A comparison with H_A is not necessary since the constraints of this hypothesis are not supported by the data anyway. In conclusion, there is no support for the expectation that an increase in hostile intentions takes place for all three groups following emotion manipulation, but there is support for the expectation that the increase in hostile intentions becomes larger when the groups consist of more aggressive boys.

Similar computations can be performed for the aggressive response; see Table 4. However, none of the hypotheses under investigation are better than an unconstrained hypothesis. Consequently, none of the hypotheses gives an adequate description of the population from which the data were sampled. As a result, there is no increase in aggressive response following emotion manipulation, and there is no support for the expectation that the increase in aggressive response becomes larger when the groups consist of more aggressive boys. A combination of both hypotheses, H_C, receives even less support.

Complexity and Fit

For a better understanding of the Bayes factor and its relation with model fit and model complexity, we elaborate on f_i and c_i. As was shown before, Bayesian model selection provides the degree of support for each hypothesis under consideration and combines model fit and model complexity. It has a close link with classical model selection criteria such as the Akaike information criterion (AIC; Akaike, 1981) and the Bayesian information criterion (BIC; Schwarz, 1978) that also combine fit and complexity to determine the support for a particular model. However, in contrast with Bayesian model selection, these classical criteria are as of yet unable to deal with hypotheses specified using inequality constraints (Mulder et al., 2009; Van de Schoot, Romeijn & Hoijtink, 2009). In the specific application of Bayesian model selection used in this article, the Bayes factor's selection criteria also combine model fit and complexity, but it is able to account for inequality constraints. As we now illustrate, complexity and fit are (although implicitly) also important parts of Bayesian model selection.

Table 4
Estimates for Bayes Factors (BFs) Against H_{unc}, Model Fit, and Model Complexity for H_A, H_B, and H_C

	Hostile			Aggressive		
Hypothesis	f_i	c_i	BF	f_i	c_i	BF
H_A	.06	.25	0.24	.23	.25	0.92
H_B	.64	.16	4.00	.02	.16	0.12
H_C	.03	.02	1.50	$1e^{-6}$	.02	0.00
H_{unc}	1	1	1	1	1	1

Model complexity. The first component of the Bayes factor is model complexity, c_i, which can be computed before observing any data. The Bayes factor incorporates the complexity of a hypothesis by determining the number of restrictions imposed on the means. Note that model complexity is independent of the data because it is the proportion of the prior distribution in agreement with the constraints. Let us elaborate on this using our running example.

According to Sober (e.g., 2002), the simplicity of a hypothesis can be seen as an indicator of the amount of information the hypothesis provides. Classical model selection tools favor models that allow for fewer possibilities and call such models simpler. Relating this to, for example, the AIC and BIC, where complexity is measured as the number of parameters in a model, the more dimensions are "shaved" away, the simpler the model becomes. We maintain that there is also such a natural relation between introducing inequality constraints and ruling out possibilities, that is, when specifying such inequality constraints, a researcher also "shaves" away parameter space volume. In sum, a simple hypothesis contains more restrictions and more information and as such, is more specific and should be favored by the model selection procedure.

Returning to our example, the most complex hypothesis is always H_{unc}, in the sense that all combinations of means are allowed and no constraints are imposed. Therefore, c_i for H_{unc} is equal to 1; see Table 4. Let us consider the hypotheses specified for hostile attribution. There are two constraints specified for H_{B} (see Table 1). Consequently, not all combinations of means are possible. H_{B} is therefore considered simpler than H_{unc}. Three constraints are specified for H_{A}, and this hypothesis is even simpler than H_{B}. The simplest hypothesis is H_{C}, because here the most information is added: the constraints of H_{A} in addition to the constraints of H_{B}. With respect to complexity, the hypotheses can be ordered from simplest to most complex: H_{C}, H_{B}, H_{A}.

In Table 4, estimates for model complexity are displayed and our expected ordering for both hostile attribution and aggression is confirmed. That is, the proportion of the prior distribution in agreement with the constraints for H_{A} is .25 and for H_{C} only .02, making this latter hypothesis less complex because more information is specified in term of the number of inequality constraints.

Model fit. After observing some data, the second component of the Bayes factor is model fit, f_i. Loosely formulated, it quantifies the amount of agreement of the sample means with the restrictions imposed by a hypothesis.

Consider the sample means in Table 3. The observed sample means fit perfectly with an unconstrained hypothesis because no constraints are imposed on the means. Consequently, H_{unc} always has the best model fit compared with any other informative hypothesis and $f_i = 1$; see Table 4. With respect to the informative hypothesis on hostile attribution, it appears that one constraint is violated for H_{A}: The sample mean of the nonaggressive group for the neutral condition is higher for the negative condition rather than lower. As a result, the model fit of H_{A} is worse than the model fit for H_{unc}. For H_{B} there appeared to be no violations of the constraints. Since H_{C} is a combination of the constraints of H_{A} and H_{B}, there is one violation of the constraints imposed by this hypothesis. In sum, with regard to model fit, H_{B} performs better than H_{A} and H_{C}, respectively.

In Table 4, estimates for model fit for the three informative hypotheses on hostile attribution are displayed and as can be seen, the expected ordering is confirmed. With regard to three informative hypotheses on aggression, the fit is rather low for all three hypotheses. After computing f_i and c_i, the Bayes factor shown in Equation (1) can be computed, for example, for hostile attribution:

$$\text{BF}_{\text{B,unc}} = \frac{f_i}{c_i} = \frac{.64}{.16} = 4. \tag{3}$$

As was correctly noticed by one of the reviewers, it can be illustrative to provide more information than just the Bayes factors in terms of model fit and model complexity. Information about the posterior distributions of the means and their credibility intervals can be found in Figure 1. The interpretation of a Bayesian 95% credibility interval is that, for example, the posterior probability that $M_{neu,non}$ for hostile attribution lies in the interval from $-.32$ to .66 is .95 (see, e.g., Lynch, 2007). 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 $M_{neu,non}$ for aggression is .58 and that there is a .95 probability that it is between .32 and .86. This credibility interval does not include zero and, consequently, the null hypothesis $M_{neu,non} = 0$ is rejected. Furthermore, it can be seen that the credibility intervals for $M_{neu,non}$ and $M_{neg,non}$ for aggression show an overlap, so the constraint $M_{neu,non} < M_{neg,non}$ is not supported by the data. Suppose we observed the same results (i.e., posterior means) but with a larger sample size: The posterior distributions would be more peaked. Hence, the overlap of the credibility intervals for $M_{neu,non}$ and $M_{neg,non}$ will disappear. Consequently, the fit of the model would increase. In the next section, we elaborate on the relation between model fit on one hand, and effect size and sample size on the other hand.

Bayes Factors Versus *p* Values

Recall that a Bayes factor provides a direct quantification of support as evidenced in the data for two competing hypotheses. Most researchers would agree that 100 times more support seems to be quite a lot and, for example, that 1.04 times more support is not that much. However, clear guidelines are not provided in the literature, and we do not provide these either. We refrain from doing so 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).

To gain insight into the interpretation of Bayes factors in comparison with *p* values, consider the following imaginary example. Suppose there are six means, denoted by $M_1, \ldots, M_6$, and that the informative hypothesis of interest is H_{D}: $M_1 < M_2 < M_3 < M_4 < M_5 < M_6$. We created data in such a way that the sample means and variance correspond exactly to population values, as shown in Footnote 1 in Table 5. Now let us compare

1. The F test for traditional frequentist analysis,
2. Planned comparisons for thoughtful frequentist analysis assuming a linear increase

$$(-2.5 \times M_1 + -1.5 \times M_2 + -.5 \times M_3 + .5 \times M_4 + 1.5 \times M_5 + 2.5 \times M_6),$$

and

3. Bayesian evaluation of informative hypotheses using BFs as described above for $\text{BF}_{\text{D,unc}}$.

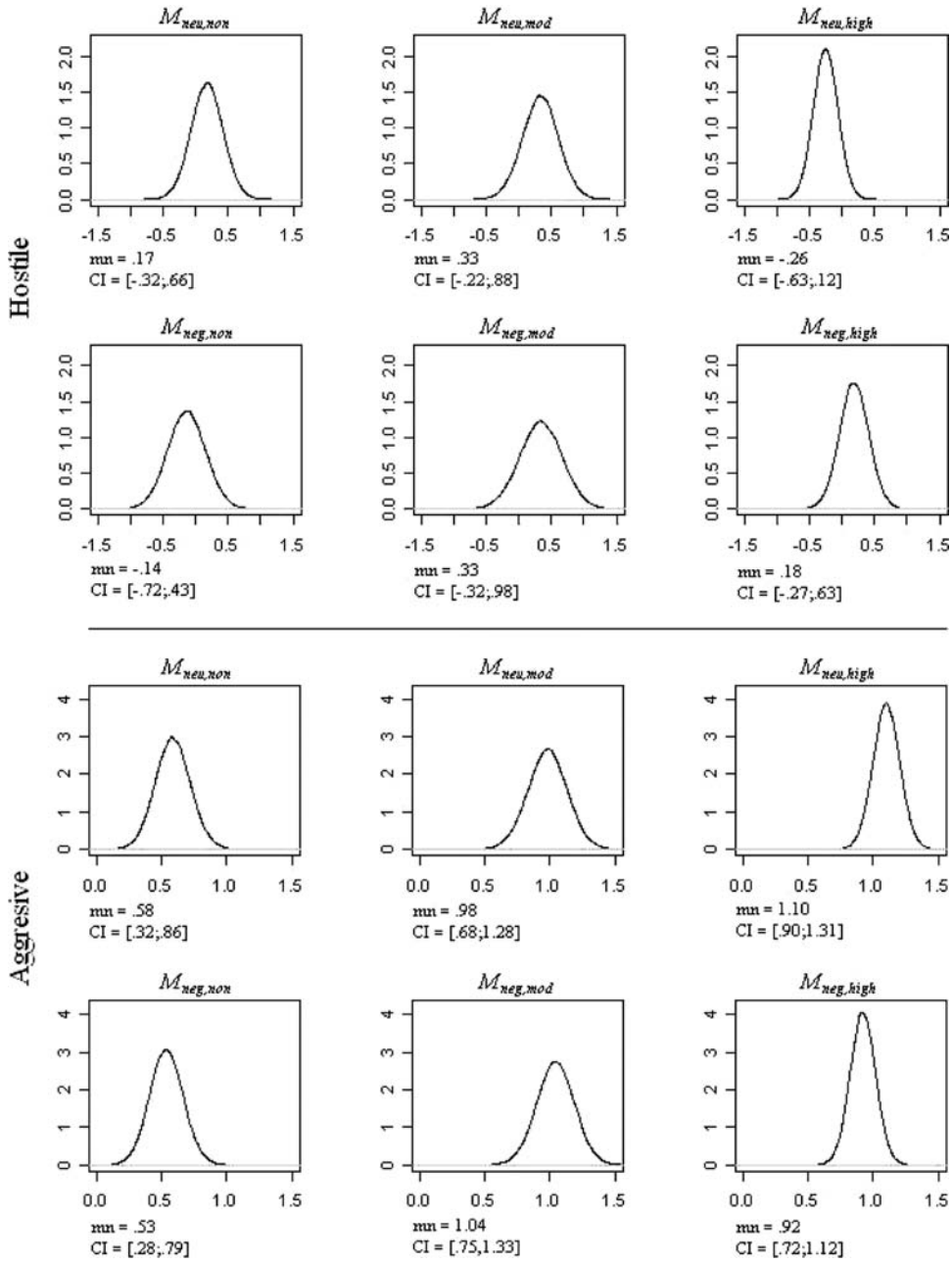

Figure 1. Posterior distributions for all groups on the dependent variables *hostile attribution* and *aggressive responses*. Note that "mn" denotes posterior mean and "CI" denotes the Bayesian credibility interval.

We ran these analyses for different populations with a small and medium effect, a small and large sample size, and zero, one, and two violations of the ordering; see Table 5. Comparison of the resulting p values with the Bayes factors provides insight into the interpretation.

As can be seen in Table 5, for some of the data, the classical F test is not significant, although there are differences between the means within the population (i.e., Populations 2, 6, 8). This result indicates a power problem that is not shared by the planned comparison and the Bayes factor. The results for the planned comparison indicate that for all populations, apart from the null Population 1, there is a significant linear increase in the six means even with one or two violations of the constraints.

Inspection of the Bayes factors indicates that their value is dependent on, first, effect size. Compare, for example, Population 2 with Population 4, with Bayes factors 29 versus 91, respectively. Second, consider the sample size and compare, for example, Population 2 with Population 3, with Bayes factors 29 versus 470, respectively. Finally, look at the number of violations. Compare, for example, Populations 2, 6, and 8 with 0, 1, and 2 violations and with Bayes factors of 29,

Table 5
Results of the Comparison Between a Classical F Test, Planned Comparison, and Bayes Factors

Population	Small/medium effect[a]	Small/large sample size per group	0/1/2 violations[b]	Classical *F* test	Linear increase	Bayes factor (BF) versus unconstrained model
1	No effect[c]	100	0	$F = 0$; $p = 1$	Contrast = 0; $p = 1/2$	BF = 1.05
2	Small	10	0	$F = 1.40$; $p = .15$	Contrast = 0.84; $p = .01/2$	BF = 29.51
3	Small	100	0	$F = 14.0$; $p < .001$	Contrast = 0.84; $p < .001$	BF = 470.29
4	Medium	10	0	$F = 3.58$; $p = .007$	Contrast = 1.34 $p < .001$	BF = 91.55
5	Medium	100	0	$F = 35.84$; $p < .001$	Contrast = 1.34; $p < .001$	BF = 694.27
6	Small	10	1	$F = 1.40$; $p = .24$	Contrast = 0.79; $p = .02/2$	BF = 20.52
7	Small	100	1	$F = 14.0$; $p < .001$	Contrast = 0.79; $p < .001$	BF = 48.22
8	Small	10	2	$F = 1.40$; $p = .23$	Contrast = 0.74; $p = .02/2$	BF = 12.74
9	Small	100	2	$F = 14.0$; $p < .001$	Contrast = 0.74; $p < .001$	BF = 4.75

[a] Effect size according to definition of Cohen (1992) with population means for the small effect: −.50, −.30, −.10, .10, .30, .50 (*SD* = 1; effect size = .11), and for the medium effect: −.80, −.48, −.16, .16, .48, .80 (SD = 1; effect size = .28). [b] With 1 violation two means are reversed (e.g., −.50, −.10, −.30, .10, .30,.50) and with 2 violations four means are reversed (e.g. −.50, −.10,−.30, .10, .50, .30). [c] All means are zero in the population.

20, and 4, respectively. In the latter population there is still support for the informative hypothesis, but 4 is clearly not a great deal of support in comparison to the other, much larger, results.

Recall the posterior means presented in Figure 1. Suppose the sample size is increased; then the posterior distributions will become more peaked, and the overlap between distributions will disappear. Stated otherwise, the Bayes factor will increase with increasing sample size because of an increase in model fit. The same holds for an increase in effect size, that is, the further the posterior means are away from each other, the less overlap there is between distributions.

What can be learned from this exercise? First, Bayes factors are sensitive for effect size, the number of violations, and sample size. When comparing informative hypotheses, the complexity of each hypothesis under investigation is independent of these three concepts, as we showed before. It is the fit of the model that is influenced by the three concepts; in other words, the fit of a model will increase with higher effect sizes, a decrease in the number of violations, and an increase in sample size. Second, in this section we specified only one single informative hypothesis, which we evaluated with Bayes factors and *p* values. It is interesting to note that the Bayes factor tells us exactly how much better a certain informative hypothesis is against another hypothesis. In comparison, a *p* value tells us the probability, given that the null hypothesis is true, of observing the same data or more extreme data than those actually observed. The *p* value, however, is often misinterpreted as the probability that the null hypothesis is true. Recall that if we were to specify more informative hypotheses, it would be difficult, or even impossible, to use *p* values as was shown before.

Conclusion

In this article, we showed how subjective beliefs influence analyses in hidden ways and how they might be incorporated explicitly. Researchers in developmental psychology often have explicit expectations about their research questions, or as Lee and Pope (2006) say, "In the real-world much is usually already known about a problem before data are collected or observed." As we have shown, these expectations can be translated into informative hypotheses. However, as we demonstrated with a case study, the average researcher wants to evaluate such informative hypotheses but tests a set of null hypotheses. We argued that researchers should not use traditional frequentist analysis, not even thoughtful frequentist analysis, if they are not interested in the conclusion that the observed data either are or are not in agreement with the null hypothesis. Rather, researchers should directly evaluate all of the informative hypotheses under investigation without relying on a test of the null hypothesis. This can be done using Bayesian model selection. This way, researchers can use all of the knowledge available from previous investigations and can learn more from their data than they can with traditional null hypotheses testing. All criticisms of null hypothesis testing aside, the best argument for evaluating informative hypotheses is probably that, like Orobio de Castro et al. (2003), many researchers want to evaluate a set of hypotheses formulated with inequality constraints but have been unable to do so because the statistical tools were not yet available. As this article has illustrated, these tools are available to any researcher.

References

Achenbach, T. M. (1991). *Manual for the Teacher's Report Form and 1991 profiles.* Burlington: University of Vermont, Department of Psychiatry.

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289–300.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Coie, J. D., & Dodge, K. A. (1998). Aggression and antisocial behavior. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology* (5th ed., Vol. 3, pp. 779–862). Toronto, Canada: Wiley.

Deković, M., Wissink, I., & Meijer, A. M. (2004). The role of family and peer relations in adolescent antisocial behaviour: Comparison of four ethnic groups. *Journal of Adolescence, 27,* 497–514.

Dodge, K. A. (1985). Attributional bias in aggressive children. In P. C. Kendall (Ed.), *Advances in cognitive-behavioral research and therapy* (pp. 73–110). Orlando, FL: Academic.

Gelman, A.,Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd edition). London: Chapman & Hall.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6,* 65–70.

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.

Hsu, J. C. (1996). *Multiple comparisons.* London: Chapman and Hall.

Kammers, M., Mulder, J., De Vignemont, F., & Dijkerman, H. (2009). The weight of representing the body: Addressing the potentially indefinite number of body representations in healthy individuals. *Experimental Brain Research, 204,* 333–342.

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

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

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods, 15,* 281–299.

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory 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: Erlbaum.

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

Lee, M. D., & Pope, K. J. (2006). Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference. *Journal of Mathematical Psychology, 50,* 193–202.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112,* 662–668.

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

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147–163.

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. (2010). *Change and stability of personality types in adolescence: A five-wave longitudinal study in early-to-middle and middle-to-late adolescence.* Manuscript submitted for publication.

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

Orobio de Castro, B., Slot, N. W., Bosch, J. D., Koops, W., & Veerman, J. W. (2003). Negative feelings exacerbate hostile attributions of intent in highly aggressive boys. *Journal of Clinical Child and Adolescent Psychology, 32,* 56–65.

Orobio de Castro, B., Veerman, J. W., Koops, W., Bosch, J. D., & Monshouwer, H. J. (2002). Hostile attribution of intent and aggressive behavior: A meta-analysis. *Child Development, 73,* 916–934.

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

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

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology, 43,* 1124–1139.

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

Sober, E. (2002). *Bayesianism, its scope and limits.* Oxford: Oxford University Press.

Van de Schoot, R., Hoijtink, H., & Doosje, S. (2009). Rechtstreeks verwachtingen evalueren of de nul hypothese toetsen? nul hypothese toetsing versus bayesiaanse model selectie [Directly evaluating expectations or testing the null hypothesis: Null hypothesis testing versus Bayesian model selection]. *De Psycholoog, 4,* 196–203.

Van de Schoot, R., Romeijn, J.-W., & Hoijtink, H. (2010). *Background knowledge in model selection procedures.* Manuscript submitted for publication.

Van de Schoot, R., & Wong, T. (in press). Do antisocial young adults have a high or a low level of self-concept? *Self and Identity.*

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 match and mismatch effects. *Behavior Modification, 32,* 427–449.

Verhulst, F. C., van der Ende, J., & Koot, H. M. (1997). *Handleiding voor de Teacher's Report Form (TRF): Nederlandse versie* [Manual for the Teacher's Report Form (TRF): Dutch version]. Rotterdam, The Netherlands: Erasmus Universiteit Rotterdam.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* -values. *Psychonomic Bulletin & Review, 14,* 779–804.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 4,* 212–213.

Walker, L. J., Gustafson, P., & Frimer, J. A. (2007). The application of Bayesian analysis to issues in developmental research. *International Journal of Behavioral Development, 31,* 366–373.

Walker, L. J., Gustafson, P., & Hennig, K. H. (2001). The consolidation/transition model in moral reasoning development. *Developmental Psychology, 37,* 187–197.

Received September 9, 2008
Revision received July 5, 2010
Accepted July 8, 2010 ■