Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection

[Under review for Developmental Psychology]

Rens van de Schoot, Herbert Hoijtink and Joris Mulder
Department of Methodology and Statistics
Utrecht University, The Netherlands

Marcel A. G. Van Aken and Bram Orobio de Castro
Department of Developmental Psychology
Utrecht University, The Netherlands

Wim Meeus
Department of Child and Adolescent Studies
Utrecht University, The Netherlands

Jan-Willem Romeijn
Department of Philosophy
Groningen University, The Netherlands

Correspondence should be addressed to Rens van de Schoot: Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3508TC, Utrecht, The Netherlands; Tel.: +31 302534468; Fax: +31 2535797; E-mail address: a.g.j.vandeschoot@uu.nl

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

Abstract

Researchers often have expectations about the research outcomes in regard to inequality constraints between 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 towards other peers than moderately aggressive boys, who in turn score *higher* than non-aggressive boys. In most cases, null hypothesis testing is used to evaluate such hypotheses. We will show, however, that hypotheses formulated using inequality constraints between the group means cannot be evaluated properly by means of classical null hypothesis testing nor by using one-sided hypotheses tests together with planned comparisons. These analyses test the wrong hypotheses and the testing procedure itself suffers from complications. In this paper, 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

Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection

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, or there are no group differences, etc. We argue that many researchers have expectations about various components of the analysis, for instance the ordering of means, and are not particularly interested in testing a classical null hypothesis (see also Cohen 1990, 1994). 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 non-aggressive boys. This expectation is clearly not the same as the classical null hypothesis: all scores for the boys are equal. We refer to such expectations as informative hypotheses.

In this paper we describe, by means of a case study, what can happen if a researcher has such informative hypotheses, and uses either classical null hypothesis testing or one-sided hypothesis testing together with planned comparisons. Subsequently, we will elaborate on alternative strategies. We briefly highlight one alternative: possibilities in the field of structural equation modelling, in particular Bayesian model selection. Furthermore, we use one of our own studies 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, Slot, Bosch, Koops, and Veerman (2003) investigated the effects of inducing a negative emotional state in aggressive boys. It was questioned whether inducing

negative emotions would make boys with aggressive behavior problems attribute more aggressive responses and hostile intentions to their peers in comparison to the group of non-aggressive boys. The authors examined three levels of aggression: high, moderate, and no aggression. Mild negative emotions were induced 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*). Hostile intent attributions and aggressive responses to other peers were assessed 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: (1) why the provocateur in the vignette acted the way he did; (2) how the participants would respond if they were 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. By counting the number of vignettes in each condition with a hostile or an aggressive response to the questions, respective scores for hostile intentions and responses were calculated.

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 based on Dodge (1985), who hypothesized that a negative emotional state makes children more

prone to attribute hostile intentions to other children they interact with. The constraints corresponding to the informative hypothesis $H_{A,host}$ in relation to hostile attribution are displayed in Table 1. It can be seen, for example, that the mean score for non-aggressive boys in the neutral condition is expected to be lower than the mean score for non-aggressive boys in the negative condition, $M_{neu,\ non} < M_{neg,\ non}$. Note that the same constraints hold for aggressive responses ($H_{A,aggr}$).

A second expectation (B) was that emotion manipulation would influence aggressive boys more than 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 non-aggressive 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 non-aggressive boys do (see also, Orobio de Castro, Veerman, Koops, Bosch, & Monshouwer, 2002). The constraints corresponding to the informative hypothesis for hostile attribution ($H_{B,host}$) 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 non-aggressive 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 ($H_{B,aggr}$).

A third expectation (C) was a combination of expectation 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 become larger if boys are

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 investigated throughout the current paper is which of these three informative hypotheses, H_A, H_B, or $H_{C,}$ are best supported by the data. We try to answer this research question using classical null hypothesis testing, one-sided hypothesis testing, planned comparisons, structural equation modelling and finally Bayesian model selection. The latter procedure will be more thoroughly introduced below.

Results from the Null Hypothesis Testing

An often used strategy to analyse data like ours is classical null hypothesis testing. In our example, aggressive responses and hostile intentions were used as dependent variables in two 3 × 2 analyses of variance with *level of aggression* (high, moderate, and no aggression) as a between-participants factor and the *condition* (neutral/negative) as a within-participants factor. Three hypotheses were tested for both hostile intentions and aggressive responses:

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

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

$H_{0,1x2}$: There is no interaction between levels of aggression and the condition.

From the results, it can be seen that the only significant interaction is found for hostile attribution *(level of aggression* × *condition*); see Table 2. There were no differences between condition means for both aggressive responses and hostile intentions. However, for both aggressive responses and hostile intentions there appear to be significant differences between aggression level means (see Table 2).

Many researchers would now perform a follow-up analysis, which we also doin the section entitled ‘The Evaluation of Informative Hypotheses’. However, we first show what happens if the informative hypotheses H_A, H_B, and H_C are analysed using the null hypotheses $H_{0,1}$ $H_{0,2}$ and $H_{0,\ 1x2}$.

What Can Happen?

Although classical null hypothesis testing has been the dominant research tool for the latter half of the past century it suffers from serious complications (e.g., Wagenmakers, 2007), particularly when evaluating informative hypotheses like H_A, H_B, and H_C. In this section, we discuss what can happen when the null hypotheses $H_{0,1}$ $H_{0,2}$ and $H_{0,\ 1x2}$ are tested to answer the question which informative hypothesis (H_A, H_B, or H_C) is best supported by the data.

The Null and Alternative Hypothesis

The first problem that arises is that there is no straightforward relationship between the informative hypotheses under investigation, H_A, H_B, and H_C, and the null hypotheses that are actually being tested. De Castro et al. (2003) were not interested in testing the hypotheses $H_{0,1}$ $H_{0,2}$ and $H_{0,\ 1x2}$ that were tested in the previous section. 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 also 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 $H_{0,\ 1x2}$'. As can be seen in Table 2, some of the null hypotheses are rejected in favor of the alternative hypothesis (significant results 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 these 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 classical 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. Inspecting Table 3, it appears there is a violation of expectation A with regard to hostile attribution: the mean of the non-aggressive group is lower in the negative condition

than 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 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 hypothesis B. However, does this imply that H_B is preferred over H_A? What if there would have been a small deviance of the constraints imposed on the mean differences: -.45, -.46, .45? Or what if there would have been a larger deviance between the mean differences: -.45, -.55, .45? When would the difference be large enough to conclude that the informative hypothesis holds?

Multiple Hypothesis Testing and Power

Alongside the complication of testing the wrong hypotheses, the procedure of classical null hypothesis testing suffers from a number of complications itself. Two important issues will be discussed 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. Multiple testing increases the family-wise error rate, which is the probability of incorrectly rejecting at least one null hypothesis of all hypotheses tested. For example, for two independent tests and an alpha level of .05 per test, the probability of correctly concluding that both H_0's are true is .95 x .95 = .90 and for six tests this is $.95^6 = .74$. In the latter case, the probability of incorrectly rejecting at least one null hypothesis is 1 -.74 = .26. Note that the six tests in Table 2 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 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). If the alpha level is corrected, this also requires a larger

sample size to maintain sufficient power, which may not always be realistic. In this 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 of 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 Hsu1996). However, larger sample sizes are still needed in these cases, and for it remain difficult to determine how the overall alpha level should be corrected with all of these methods. For example, when using the Bonferroni correction, should the overall alpha be corrected separately for each dependent variable, so α/3? Or should the overall alpha be corrected by using the total number of tests, so α/6? The answers to these questions are not clear and similar complications hold for the false discovery rate and the Holm-Bonferroni method.

If we were to use the Bonferroni correction (α/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, or H_C? For aggressive responses, aggression level differences remain significant when using α/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) ≠ (1.02 + 1.08 = 1.10) ≠ (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.

The Evaluation of Informative Hypotheses

What have we learned so far? Testing the null hypotheses $H_{0,1}$ $H_{0,2}$ and $H_{0,\ 1x2}$ 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 expectations in the form of inequality constraints between means, he or she might be better off by using alternative procedures. In this section, we use a combination of one-sided hypothesis testing and planned comparisons to evaluate H_A, H_B, and H_C. We then take a side trip to structural equation modelling. Structural equation modelling is a flexible tool that can deal with many types of constraints, making it a useful tool in this situation. Finally, we will present a Bayesian method that is, as of yet, the only method that allows a direct evaluation of H_A, H_B, and H_C.

One-sided Hypothesis Testing and Planned Comparisons

If two means (or difference between means) are ordered, hypothesis testing can be made directional by dividing the p-value for the corresponding test by 2. In our example of Orobio de Castro et al. (2003), two sets (one for each dependent variable) of three one-sided t-tests can be performed for 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, three difference scores can be computed [$M_{neu,non}$ - $M_{neg,non}$], [$M_{neu,mod}$ - $M_{neg,mod}$], and [$M_{neu,high}$ - $M_{neg,high}$]. An approximation of H_B can be obtained by using a linear contrast built with these scores. A good primer of using planned comparisons is presented in Rosenthal, Rosnow, and Rubin (2000), where several types of contrasts are introduced. In our example, H_B can be evaluated using the linear contrast -1 x [$M_{neu,non}$ - $M_{neg,non}$] + 0 x [$M_{neu,mod}$ - $M_{neg,mod}$] + 1 x [$M_{neu,high}$ - $M_{neg,high}$]. Since this hypothesis is also directional, we expect an increase in the difference between conditions; the resulting *p*-value can be divided by two

($p_{hostile}$ = .008/2; p_{aggr} = .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.

Although the above procedure generates better results than the naïve procedure presented in the previous section, there are still some problems related to one-sided hypothesis testing and planned comparisons. Recall that we wanted to evaluate H_A, H_B and H_C. One-sided hypothesis testing and planned comparisons results in four *p*-values per dependent variable. The first concern is that once again, null hypotheses are being tested, for example $M_{neu,non} = M_{neg,non}$ versus $M_{neu,non} < M_{neg,non}$. Although these tests are an improvement over the hypothesis testing shown in the previous section, these tests are not the same as evaluating H_A versus H_B versus H_C. Suppose all null hypotheses are rejected, what information do we now have to evaluate H_A, H_B, and H_C?

Secondly, what do we do with contradictory results? For instance, there are no significant differences between $M_{neu,non}$ and $M_{neg,non}$ and between $M_{neu,mod}$ and $M_{neg,mod}$, but there is a significant difference between $M_{neu,\ high}$ and $M_{neg,high}$. Should H_A be rejected? What if the *p*-values were .051, .051 and .051, respectively? These issues arise because once again, the wrong hypotheses are being tested.

Thirdly, there is still the issue of multiple hypothesis testing: there are four p-values per dependent variable. Should we divide all *p*-values by four or by eight? A researcher must decide which *p*-values to include in the family of hypotheses which they are attempting to correct.

Finally, the planned contrast assumes a linear relationship, whereas H_b only assumes a monotone relationship: $[M_{neu,non} - M_{neg,non}] < [M_{neu,mod} - M_{neg,mod}] < [M_{neu,high} - M_{neg,high}]$. Although the linear contrast is an approximation of this monotone relationship, the assumption of linearity is maybe too much to ask for in this example. It could be argued that it is more likely that there is a large difference between non-aggressive boys versus aggressive boys, but

that the difference between moderate and highly aggressive boys is much smaller. In sum, although an improvement over the evaluation of main and interaction effects, even one-sided testing and the evaluation of well chosen contrasts are not perfectly suited for the evaluation of H_A versus H_B versus H_C.

Structural Equation Modelling

Structural equation modelling can be a useful tool for evaluating a set of competitive hypotheses. Researchers who fit structural equation models can impose many simultaneous hypothesized relationships between variables. Moreover, various constraints such as nonlinear, equality, and inequality constraints can be imposed upon the model parameters.

When using structural equation modelling, researchers often specify a set of structural models to evaluate their hypotheses, for example by constraining regression coefficients to be equal between groups in one model, and relaxing these constraints in a second model. This approach works well when equality constraints are used, however, when inequality constraints are used, comparisons between structural equation models become problematic. Problems arise because default model comparison tools, for example AIC (Akaike, 1981) and BIC (Schwartz, 1978), are not equipped to deal with inequality constraints (Mulder et al. 2009b). It is currently not known how to quantify the number of parameters if these are restricted using inequality constraints. Therefore, the penalty term is undetermined in this case and AIC and BIC cannot be used as a comparison tool between inequality constraint models.

Applicable tools for evaluating inequality constraint hypotheses in structural equation modelling need to be developed further (Gonzalez & Griffin, 2001; Stoel, Galindo-Garre, Dolan, & Van den Wittenboer, 2006; Van de Schoot, Hoijtink & Deković, in press). Note, however, that our example of Orobio de Castro et al. (2003) is not a structural equation model and as we will show below, tools for evaluating analysis of variance models are available. Generalisation of these procedures is feasible but as of yet, are not available.

Bayesian Methods

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 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 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 study of Schulz, Bonawitz, and Griffiths (2007) about causal learning processes in pre-schoolers. 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.

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, Klugkist, &, Boelen, 2008; Klugkist, Laudy, &, Hoijtink, 2005, Kuiper & Hoijtink, 2009; Laudy, Boom, & Hoijtink, 2005; Laudy & Hoijtink, 2007; Mulder, Hoijtink, & Klugkist, 2009a; Mulder et al., 2009b). 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, Kolk, and Klugkist (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 matches 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, Van de Schoot, Keijsers, Schwartz, and Branje (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 paper 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 is directly evaluated, there is no

need to deal with contradictory results or problems arising as a result of multiple testing.

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, and we refer interested readers to Gelman, Carlin, Stern, and Rubin (1995) for a general introduction to Bayesian analysis. For incorporating inequality constraints in the context of Bayesian model selection, we refer interested readers to Hoijtink et al. (2008).

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.

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 criterion such as AIC (Akaike, 1981) and BIC (Schwartz, 1978) that also combine fit and complexity to determine the support for a particular model. However, in contrast to Bayesian model selection these classical criteria are as of yet unable to deal with hypotheses specified using inequality constraints (Mulder et al., 2009b). In the specific application of Bayesian model selection used in this paper, the Bayes factors selection criteria also combine model fit and complexity, but are able to account for inequality constraints. Note,

however, that the interpretation of fit and complexity is somewhat different than the interpretation of fit and complexity when using AIC and BIC. Let us elaborate on this.

The first component of the Bayes factor is model fit. 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 to any other informative hypothesis. With respect to the informative hypothesis on hostile attribution, it appears that one constraint is violated for H_A: the sample mean of the non-aggressive 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 appear to be no violations of the constraints; consequently, this hypothesis has the same model fit as H_{unc}. 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_{unc} and H_B perform better than H_A and H_C.

The second component of the Bayes factor is complexity. According to Sober (2002), the simplicity of a hypothesis can be seen as an indicator of the amount of information the hypothesis provides. A simple hypothesis contains more restrictions and contains more information and as such, is more specific. In other words, the more information a researcher is able to add to a hypothesis using inequality constraints, the simpler it becomes. Loosely formulated, the Bayes factor incorporates the complexity of a hypothesis by determining the number of restrictions imposed on the means.

The most complex hypothesis is always H_{unc}, in the sense that all combinations of means are allowed and no constraints are imposed. 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, H_{unc}.

Bayes factors combine model fit and complexity and represent 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 4 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.

In this paper we analysed the informative hypotheses of our example using the software presented in Mulder et al. (2009a). The method described in Mulder et al. can deal with many complex types of (in)equality constraints in multivariate linear models, e.g. MANCOVA, regression analysis, 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). It was 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 latent class analysis (Hoijtink, 1998, 2001; Laudy et al, 2005) as well as order restricted contingency tables (Laudy & Hoijtink, 2007). Readers interested in this software can visit *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 analysing data can best be made by using the computations executed in the software

program ‘confirmatory ANOVA’ (Kuiper, Klugkist, & Hoijtink, 2009). We refer to Klugkist et al. (2005), Mulder et al., (2009a, 2009b), Laudy et al, (2005) and Laudy & Hoijtink, (2007) for technical details on actual computations.

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, 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 to *p*-values, consider the following imaginary example. Suppose there are six means and that the informative hypothesis of interest is H: $M_1 < M_2 < M_3 < M_4 < M_5 < M_6$. The data were generated in such a way that the sample means and variance correspond to the population values (see Table 4). We computed the F-test (classical null hypothesis), planned comparisons (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 Bayes factors (monotone increase: $M_1 < M_2 < M_3 < M_4 < M_5 < M_6$) for different populations (small/medium effect, small/large sample size, 0/1/2 violations of the ordering; see Table 4).

As can be seen in Table 4, for some of the hypotheses the classical F-test is not significant (i.e. population 2, 6, 8), although there are differences between the means within the population. 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 1 or 2 violations of the constraints. Inspection of the Bayes factors indicates that its value is

dependent on: (i) effect size: compare for example population 2 with population 4, with Bayes factors 32 versus 156, respectively; (ii) sample size: compare for example population 2 with population 3, with Bayes factors 32 versus 393, respectively; (iii) the number of violations: compare for example populations 2, 6 and 8 with 0, 1 and 2 violations and with Bayes factors of 32, 6 and 1.94, respectively. In the latter population there is still support for the informative hypothesis, but 1.94 is clearly not a great deal of support in comparison to the other, much larger, results.

Note that in this example we only formulated a simple ordered hypothesis that can easily be approximated by one single contrast. However, for a more complex informative hypothesis, more planned comparisons are needed and as a result, more *p*-values need to be inspected. In contrast, the result of Bayesian model selection always consists of one value for each pair of informative hypotheses.

Example Reconsidered

To analyse the data of Orobio de Castro et al. (2003) we computed the Bayes factors using two analysis of variance models including a within variable (condition) and a between variable (aggression level). The results for our example are presented in Table 5.

For hostile attribution the $BF_{A,unc}$ of H_A compared to H_{unc} is 0.27. 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 to H_{unc} is 3.90, indicating that support from the data is 3.90 times stronger for H_B than for H_{unc}. The $BF_{C,unc}$ indicates that support from the data is 1.39 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 BF for each informative hypothesis with the unconstrained hypothesis (Klugkist et al., 2005). For example, the $BF_{A,B}$

between hypothesis H_A and H_B is $\frac{3.90}{0.27} = 14.44$, indicating that the support for H_B is 14.44 times stronger than the support for H_A. The $BF_{B,C}$ between H_B and H_C is $\frac{3.90}{1.39} = 2.81$, which means that the support for H_B is 2.81 times stronger than the support for H_C.

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 5. However, none of the hypotheses under investigation are better than an unconstrained hypothesis. Consequently, none of the hypotheses give an adequate description of the population from which the data were sampled. In conclusion, there is no increase in aggressive response following emotion manipulation. There is also 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.

As was correctly noticed by one of the reviewers, it can be illustrative to provide more information than just the Bayes factors. 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 lies in the interval from -.32 to .66 is 0.95 (see, e.g., Gelman et al. 2004). 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 there is a .95 probability that it is between .32 and .86. This credibility interval does not include zero and consequently the probability that $M_{neu,\ non} = 0$ is very low. 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 very probable.

Conclusion

Researchers in developmental psychology often have 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" (see also Walker et al., 2007). Using Bayesian model selection, researchers can use all the knowledge available from previous investigations and can learn more from their data using informative hypotheses rather than traditional null and alterative hypotheses. Although Frick (1996) and Wainer (1999) argue that there are situations where null hypothesis testing is appropriate, we argue that researchers should not be satisfied with the conclusion that the observed data either are or are not in agreement with the null hypothesis.

We have shown that Bayesian model selection is suited for the evaluation of informative hypotheses and results in a direct quantification of the support available in the data for each hypothesis under investigation. 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 paper has illustrated, these tools are now available to any researcher within the social sciences.

References

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.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.

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

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modelling: Every "One" matters. Psychological Methods, 6, 258 - 269.

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 in psychology*. 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.* Chapman and Hall, London.

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. (2009). Comparisons of means using confirmatory and exploratory approaches. *Manuscript submitted for publication.*

Kuiper, R. M., Klugkist, I., & Hoijtink, H. (2009). A Fortran 90 program for confirmatory analysis of variance. *Manuscript submitted for publication.*

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 (p. 63-82). Erlbaum: Londen.

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.

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods, 9,* 147-163.

Meeus, W., R. Van de Schoot, L. Keijsers, S. J. Schwartz, and S. Branje. (in press). 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.*

Mulder, J., Hoijtink, H., & Klugkist, I. (2009a). Equality and Inequality Constrained Multivariate Linear Models: Objective Model Selection Using Constrained Posterior Priors. *Journal of Statistical Planning and Inference.*

Mulder, J., Klugkist, I., Van de Schoot, R., W. Meeus, Selfhout, M., & Hoijtink, H. (2009b). Informative Hypotheses for Repeated Measurements: A Bayesian Approach. *Journal of Mathematical Psychology*

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, 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.

Schultz, 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*, 461-465.

Sober, E. (2002). *Bayesianism, its scope and limits,* pp. 21- 38. Oxford: Oxford University Press.

Stoel, R. D., Galindo-Garre, F., Dolan, C., & Van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 4*, 439 - 455.

Van de Schoot, R., Hoijtink, H. & Deković, M. (in press). Testing Inequality Constrained Hypotheses in SEM Models. *Structural Equation Modeling: A multidisciplinary journal.*

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. (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, 4,* 366-373.

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

Table 1

Constraints for Hypotheses A, B and C for Hostile Attribution

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

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.

Table 2

Results of the Two 3x2 Univariate Analyses of Variance

	Hostile		Aggressive	
	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

Table 3

Emotion Ratings by Aggression Level and Condition

		Hostile					Aggressive				
	Condition	No Agression		Moderate		High	No Agression		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

Table 4

Results of the Comparison between a Classical F-Test, Planned Comparison, and Bayes Factors

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

[1] Effect size according to definition of Cohen (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)

[2] 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**).

[3] All means are zero in the population

Table 5

Bayes Factors of H_A, H_B, and H_C against the Unconstrained Hypothesis H_{unc}.

	Hostile	Aggressive
H_A	0.92	0.27
H_B	0.14	3.90
H_c	0.00	1.39
H_{unc}	1	1

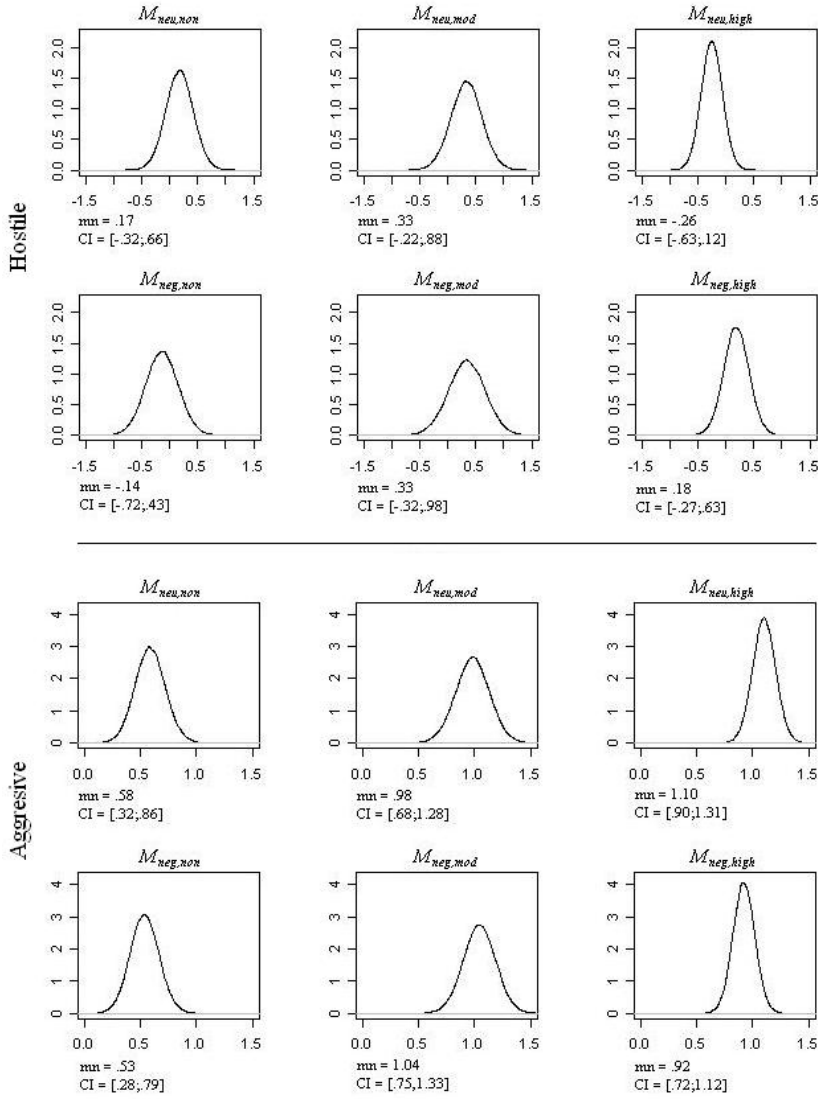

Figure 1. Posterior distributions for all groups on the dependent variables hostile attribution and aggressive responses. Note that 'mn' denotes posterior mean and 'C.I.' denotes the Bayesian credibility interval.