British Journal of Mathematical and Statistical Psychology (2015), 68, 220–245
© 2014 The British Psychological Society

The British Psychological Society

www.wileyonlinelibrary.com

Properties of hypothesis testing techniques and (Bayesian) model selection for exploration-based and theory-based (order-restricted) hypotheses

Rebecca M. Kuiper*, Tim Nederhoff and Irene Klugkist

[1]Department of Methodology and Statistics, Utrecht University, The Netherlands

In this paper, the performance of six types of techniques for comparisons of means is examined. These six emerge from the distinction between the method employed (hypothesis testing, model selection using information criteria, or Bayesian model selection) and the set of hypotheses that is investigated (a classical, exploration-based set of hypotheses containing equality constraints on the means, or a theory-based limited set of hypotheses with equality and/or order restrictions). A simulation study is conducted to examine the performance of these techniques. We demonstrate that, if one has specific, a priori specified hypotheses, confirmation (i.e., investigating theory-based hypotheses) has advantages over exploration (i.e., examining all possible equality-constrained hypotheses). Furthermore, examining reasonable order-restricted hypotheses has more power to detect the true effect/non-null hypothesis than evaluating only equality restrictions. Additionally, when investigating more than one theory-based hypothesis, model selection is preferred over hypothesis testing. Because of the first two results, we further examine the techniques that are able to evaluate order restrictions in a confirmatory fashion by examining their performance when the homogeneity of variance assumption is violated. Results show that the techniques are robust to heterogeneity when the sample sizes are equal. When the sample sizes are unequal, the performance is affected by heterogeneity. The size and direction of the deviations from the baseline, where there is no heterogeneity, depend on the effect size (of the means) and on the trend in the group variances with respect to the ordering of the group sizes. Importantly, the deviations are less pronounced when the group variances and sizes exhibit the same trend (e.g., are both increasing with group number).

1. Introduction

A central issue in most research is to evaluate the researcher's theory. When comparing group means, the researcher would often like to know whether these differ and, if so, which ones are different from each other. There are two approaches that can be used to address this question: exploratory and confirmatory (i.e., theory-based). In exploration, all the possible configurations of subsets of means are examined. That is, all possible pairs or subsets of means are examined to determine whether they are equal ('=') or not ('≠'). The number of possible configurations increases rapidly with an increase in the number of groups k. For example, when $k = 3$ and 5, there are 5 and 52 possible configurations (see

*Requests for reprints should be addressed to Rebecca M. Kuiper, Padualaan 14, 3584 CH Utrecht, the Netherlands (email: r.m.kuiper@uu.nl).

DOI:10.1111/bmsp.12041

Table 6), respectively. In the confirmatory approach, researchers solely evaluate their theories or expectations, provided that they can specify reasonable ones. This specification should be done before seeing or even collecting the data. We refer the reader to Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit (2012) for a more detailed discussion, also in light of replicability of results. Considering solely prespecified expectations typically results in a small set of hypotheses that may not include the classical null, *all means are equal* (H_0), and the classical alternative, *there are no restrictions* (H_A). This set will often include one or more order-restricted hypotheses, representing particular orderings of the means (e.g., the first three out of five group means are increasing with group number and the others are equal, that is, $\mu_1 < \mu_2 < \mu_3 = \mu_4 = \mu_5$), but also hypotheses without inequalities (e.g., the first three out of five group means are equal and the others are not restricted, that is, $\mu_1 = \mu_2 = \mu_3$, μ_4, μ_5). In adopting this approach, the researcher aims to confirm or refute his theory or compare a limited set of theories rather than to explore all the possibilities. In both exploration and confirmation, different methods can be used: hypothesis testing, model selection using information criteria, and Bayesian model selection (BMS). These two distinctions lead to six different types of techniques. Table 1 displays, for each of the six, at least one technique that can be used in the analysis of variance (ANOVA) model to evaluate group means. The descriptions of these techniques are summarized in the next section and accompanied with an illustration. A more detailed description is given in Kuiper and Hoijtink (2010).

Kuiper and Hoijtink (2010) show for one data set that, in ANOVA models, techniques evaluating order-restricted and possibly equality-constrained hypotheses perform better than those evaluating solely equalities. In this paper, we will quantify the performance of the three methods using the exploratory approach and the three employing the confirmatory approach, all applicable for ANOVA models, by means of simulation. Since theories often lead to order-restricted hypotheses, we select in our simulation study as confirmatory techniques those that can handle inequality constraints, that is, the $\bar{F}$ test, the order-restricted information criterion (ORIC), and posterior model probabilities (PMPs). Notably, the Akaike information criterion (AIC) can be used in both an exploratory and confirmatory way by examining solely theory-based hypotheses. However, it cannot evaluate order-restricted hypotheses, whereas the ORIC, a modified form of the AIC, can. In contrast, techniques able to evaluate order-restricted hypotheses can be employed in either a confirmatory or an exploratory mode. Hence, when necessary, we make a clear distinction between being suitable for examining order

Table 1. Examples of the six types of techniques for testing or evaluating hypotheses

Method	Exploration: Inspect all possibilities	Confirmation: Inspect theory/theories
Hypothesis testing	Equal n_i: Shaffer–Welch Fq (SWFq) test Unequal n_i: Tukey–Kramer (TK) test	$\bar{F}$ test; notably, the F test is a special case
Model selection based on information criteria	Paired-comparison information criterion (PCIC); e.g., PCIC based on the AIC (PCIC-AIC)	Order-restricted information criterion (ORIC); notably, the AIC is a special case
Bayesian model selection	Posterior model probabilities (PMPs)	Posterior model probabilities (PMPs)

Note. n_i is the number of observations for group i; AIC = Akaike information criterion.

restrictions or not. In this study, we will investigate both the effect of confirmation versus exploration and examining order-restricted hypotheses or not. In addition, we compare the performance of the three methods: hypothesis testing, model selection using information criteria, and BMS. Here, the performance of a technique is measured by the true hypothesis rate, that is, the percentage of times the correct hypothesis is chosen. It should be stressed that hypothesis testing serves a different purpose than model selection. The goal of the former is to reject the null hypothesis, whereas the goal of the latter is to select the best out of a set of hypotheses. Hence, in hypothesis testing, the null hypothesis is of more importance, while in model selection all hypotheses are equally important. Nevertheless, we do think it is valuable to examine the true hypothesis rates of the null hypothesis and other (order-restricted) hypotheses.[1] Owing to their purposes, we expect that the true hypothesis rate of the null hypothesis is the highest for hypothesis testing and that of non-null hypotheses for model selection.

Robustness of performance is also of importance. The effect of ANOVA assumptions violations on performance has been widely studied for the traditional ANOVA *F* test (e.g., Box, 1954; Schumacker & Akers, 2001), but little is known for techniques able to investigate order restrictions. To the authors' knowledge, only Wesel, Hoijtink, and Klugkist (2011) have examined this for BMS. To gain more insight into the performance of the three techniques appropriate for confirmatory evaluation of order restrictions, we also investigate, via simulation, their robustness to the violation of the homogeneity of variance assumption.

In the next section, we briefly illustrate six techniques that can be applied to comparing group means using an example based on Lucas (2003). Subsequently, the design and results of the two simulation studies are described. We end with a discussion of extensions to techniques appropriate for inequality constraints and of software available to implement order-restricted inference.

2. Description of the six techniques

The techniques for testing hypotheses and selecting models are introduced using data from Lucas (2003).

2.1. Example

Lucas (2003) reports a study with five experimental groups: (1) a group with a randomly selected male leader, (2) a group with a randomly selected female leader, (3) a group where the male team member who scores highest on the first task is selected as leader, (4) a group where the female team member who scores highest on the first task is selected as leader, and (5) a group in which female leadership is institutionalized and the female team member who scores highest on the first task is selected as leader. The institutionalization manipulation is achieved by showing the participants a film in which female leadership is

[1] We show the hypothesis rates for each of the techniques performed as done in practice. Hence, we do not adjust them to have, for example, equal true null hypothesis rates. Although this would make the comparison of the true non-null hypothesis rates between the three methods fairer, we prefer to show the hypothesis rates to be expected in practice. Notably, in this study, when comparing techniques of the same method, (1) the true non-null hypothesis rates are approximately the same or (2) the technique used in a confirmatory manner and suitable for order restrictions has not only a higher true non-null hypothesis rate but also a higher true null hypothesis rate than its exploratory counterpart. However, when comparing techniques belonging to another method based on true non-null hypothesis rates, one has to take into account the difference in true null hypothesis rates.

normal and females do well as leaders. The dependent variable is the influence of the leader derived from performance on a second task. The model of interest is an ANOVA model with $k = 5$ groups and $n_i = n = 30$ observations per group (for $i = 1, \ldots, 5$). The group means and standard deviations are shown in Table 2.

In this example, two hypotheses, designated H_1 and H_2, can be derived from theory. Both are based on the expectation that leaders appointed on the basis of their ability (groups 3 and 4) exert more influence over participants than leaders of the same sex appointed randomly (groups 1 and 2, respectively); that is, $\mu_1 < \mu_3$ and $\mu_2 < \mu_4$, where μ_i is the mean influence of group i. H_1 is based on two additional propositions derived from theory: *Women, according to the status characteristics theory, are disadvantaged relative to men in social interactions, all other things being equal* and *Institutionalizing women as leaders overcomes the influence gap between women and men*. In the context of the experiment, the first proposition leads to the expectation that female leaders (groups 2 and 4) exert less influence over the members of a group they lead than male leaders selected in the same manner (groups 1 and 3, respectively); that is, $\mu_2 < \mu_1$ and $\mu_4 < \mu_3$. These expectations yield $\mu_2 < \mu_4 < \mu_3$ and $\mu_2 < \mu_1 < \mu_3$, which will be written as $\mu_2 < \{\mu_1, \mu_4\} < \mu_3$ for ease of notation. The second proposition can be interpreted in a few ways. Our interpretation is that ‘overcome’ means that the gap is closed. Hence, it is expected that institutionalized female leaders selected on the basis of their ability (group 5) exert the same amount of influence over participants as male leaders appointed on the basis of their ability (group 3) and thus $\mu_5 = \mu_3$. These three expectations lead to the hypothesis

$$H_1 : \mu_2 < \{\mu_1, \mu_4\} < \mu_3 = \mu_5. \tag{1}$$

H_2 is based on theoretically derived propositions competing with those of H_1: *Female leaders selected on the basis of their competence* (group 4) *have less influence than male leaders selected at random* (group 1) and *Institutionalizing women as leaders has no effect*. The first proposition is represented by $\mu_4 < \mu_1$. Following the second proposition, it is expected that there is no difference between the influence of female leaders selected on the basis of their competence in the case of institutionalization (group 5) or in the normal case (group 4), that is, $\mu_5 = \mu_4$. These three expectations are represented by the hypothesis

$$H_2 : \mu_2 < \mu_5 = \mu_4 < \mu_1 < \mu_3. \tag{2}$$

Table 2. Sample group means and standard deviations of influence (Lucas, 2003)

Group		Mean influence	*SD*	*n*
1.	Randomly selected male leader	2.33	1.86	30
2.	Randomly selected female leader	1.33	1.15	30
3.	Male leader highest score	3.20	1.79	30
4.	Female leader highest score	2.23	1.45	30
5.	Female leader highest score and female leadership is institutionalized	3.23	1.50	30

Note. *SD* = standard deviation; *n* = group size.

2.2. Hypothesis testing

2.2.1. Exploration

The most common procedure to analyse the data of Lucas (2003) is to conduct an ANOVA F test followed by post hoc tests or pairwise multiple comparisons procedures when F is significant.[2] A powerful procedure, for equal group sizes, is the Shaffer–Welch Fq (SWFq) test (Ramsey, 2002; Toothaker, 1993, pp. 42–43, 48, note that the technique is called Shaffer–Ryan here). The SWFq test starts with an overall F test and is followed by testing only a selection of pairs of means which is based on the order of the sample means (and thus not on any expectation or hypothesis about their order). Note that this test is by definition an exploratory hypothesis testing technique.

In the omnibus test of the main effect, $F(4, 145) = 7.57(p < .001)$. Thus, the null hypothesis that the group means are all equal is rejected and we proceed with testing a certain selection of pairs of means based on the ordering of the sample means. From Table 2, it can be seen that the ordering from high to low is as follows: group 5, group 3, group 1, group 4, group 2. We will not give the details of the procedure here, as they can be found in Kuiper and Hoijtink (2010), Ramsey (2002), and Toothaker (1993, pp. 42–43, 48). Using a nominal α level of .05, it is concluded that the means of groups 2 and 5 and those of groups 2 and 3 are significantly different; all the other pairs of means are not significantly different.

From these results, it is hard to conclude anything with respect to a specific hypothesis such as H_1 and H_2 as set out in equations (1) and (2), respectively, let alone with respect to two (or more) competing theoretical expectations. This is the case even if your hypotheses do not contain order restrictions. Moreover, the results of exploratory tests may appear logically inconsistent and thus hard to interpret. For example, when $k = 3$, it is logically impossible that H_0: $\mu_1 = \mu_2$ and H_0: $\mu_2 = \mu_3$ are true if H_0: $\mu_1 \neq \mu_3$ is true. Although it might not be directly clear, our example also presents such an inconsistency. Neither the difference between the means of groups 1 and 2 nor the difference between the means of groups 5 and 1 is significant. However, the difference between the means of groups 5 and 2 is significant, apparently in conflict with the other two results. Both problems are avoided by testing the hypotheses of interest (i.e., H_1 and H_2) directly with a confirmatory hypothesis test (able to evaluate order restrictions).

2.2.2. Confirmation

The $\bar{F}$ test (Silvapulle & Sen, 2005, pp. 25–42) is a modification of the F test such that it can test theory-based hypotheses directly. Notably, theory-based hypotheses may comprise order-restricted hypotheses (such as H_1 in equation (1) or H_m: $\mu_1 \geq \mu_2 \geq \mu_3$) as well as hypotheses without inequalities (e.g., H_m: $\mu_1 = \mu_2, \mu_3$, where the first two groups are said to be equal and the third is not restricted). One can test the classical null (H_0: $\mu_1 = \ldots = \mu_k$) against a theory-based alternative (e.g., H_1) and one can test a theory-based null (e.g., H_1) against the classical alternative (H_A: $\mu_1, \ldots, \mu_k$). When conducting solely these two tests for (say) H_1, a possible result is to favour H_0 over H_1 (by not rejecting H_0 in the first test) and to favour H_A

[2] There are post hoc tests and pairwise multiple comparisons that do not require a significant overall F test; for example, the Ryan test (Toothaker, 1993).

Table 3. Results of the five $\bar{F}$ tests for the two theory-based hypotheses

Hypotheses tested	$\bar{F}$	p value
H_0 against $\boldsymbol{H_A}$	30.27	<.001
H_0 against $\boldsymbol{H_1}$	30.26	<.001
$\boldsymbol{H_1}$ against H_A	0.01	.995
H_0 against $\boldsymbol{H_2}$	22.91	<.001
$\boldsymbol{H_2}$ against H_A	7.36	.070

Note. Bold type indicates the preferred hypothesis.

over H_1 (by rejecting H_1 in favour of H_A in the second test). In that case, another test is required to conclude whether H_0 or H_A is preferred. Therefore, we recommend testing H_0 against H_A as well.[3]

In the case of Lucas, H_0 is tested against H_A, H_1, and H_2, and both H_1 and H_2 are tested against the unconstrained hypothesis H_A. The results are presented in Table 3 for $\alpha = .05$ (without multiple testing corrections). These results reveal that H_A is preferred over H_0 and that both H_1 and H_2 are preferred over both H_0 and H_A.

Although the $\bar{F}$ test provides clearer information about our hypotheses (H_1 and H_2), there is still a drawback. The conclusions from the five tests must be combined and, therefore, the results do not always lead to a single preferred hypothesis. This is also the case in Table 3. Since no direct comparison between theory-based hypotheses is possible with the $\bar{F}$ test, nothing can be concluded with respect to H_1 versus H_2. Hence, the $\bar{F}$ test is best restricted to the case of one theory-based hypothesis.

2.3. Model selection using information criteria

2.3.1. Exploration

Familiar information criteria include the Akaike information criterion and the Bayesian information criterion (BIC). They consist of a likelihood-derived ‘fit’ part and a penalty part based on complexity. The purpose of model selection is not to reject a null hypothesis but to select the best from a set of hypotheses. When applied in a classical exploratory manner, all possible configurations of means are examined. Dayton (2003) introduces the paired-comparison information criterion (PCIC), which does not examine all possibilities but only the possible configurations based on the ordered sample means. This avoids inconsistencies (Dayton, 1998, 2003) and yields higher true hypothesis rates when not all population means are equal (Cribbie & Keselman, 2000; Dayton, 2003). For $k = 5$ groups, as in the Lucas example, there are 52 possible configurations, but only $2^{k-1} = 2^4 = 16$ based on ordered sample means. It should be stressed that the PCIC, like the SWFq test, bases the set of hypotheses on the order of the sample means and not on theory, like H_1 and H_2. Thus, by definition, the PCIC is an exploratory model selection technique. The PCIC can be applied with, for example,

[3] If there is no (a priori) interest in H_0, it suffices to test only the theory-based hypothesis against the unconstrained one. Since we want to compare the $\bar{F}$ test with its exploratory counterpart and because we want to report on the null hypothesis rate, we do include H_0 in the set and thus do three tests.

Table 4. Results of model selection (PCIC-AIC) and Bayesian model selection (PMP) for the 16 hypotheses based on ordered sample means

Ordered sample means: 1.33 2.23 2.33 3.20 3.23

Corresponding group no.: 2 4 1 3 5

Model no. (m)	Model		q_m	log L_m	$PCIC\text{-}AIC_m$	PMP_m
1	{24135}	= H_0	2	−292.27	588.54	.00
2	{2,4135}		3	−283.38	572.76	.03
3	{24,135}		3	−283.68	573.36	.03
4	{241,35}		3	−281.79	569.57	.11
5	{2413,5}		3	−288.36	582.71	.00
6	{2,4,135}		4	−281.27	570.53	.04
7	**{2,41,35}**		4	−278.08	**564.16**	**.45**
8	{2,413,5}		4	−281.54	571.09	.03
9	{24,1,35}		4	−280.57	569.14	.06
10	{24,13,5}		4	−282.84	573.67	.01
11	{241,3,5}		4	−281.78	571.57	.02
12	{2,4,1,35}		5	−278.05	566.10	.09
13	{2,4,13,5}		5	−280.39	570.79	.02
14	{2,41,3,5}		5	−278.08	566.16	.01
15	{24,1,3,5}		5	−280.57	571.13	.09
16	{2,4,1,3,5}	= H_A	6	−278.05	568.10	.02

Note. PCIC-AIC = paired-comparison information criterion based on the Akaike information criterion, with $PCIC\text{-}AIC_m = -2 \log L_m + 2\, q_m$; PMP = posterior model probability. Bold type indicates the preferred hypothesis: the lowest PCIC-AIC value in model selection and the highest PMP value in Bayesian model selection.

AIC (PCIC-AIC) or BIC (PCIC-BIC). Notably, the small-sample-corrected AIC (AICC) could also be used (if $\sum_{i=1}^{k} n_i$ is small). Since Burnham and Anderson (2002, section 6.4) argue that the AIC has theoretical advantages over the BIC and the model selection criterion able to handle order restrictions (which will be discussed next) is a modification of the AIC, we will only evaluate the PCIC-AIC.

For the example, the order of the sample means and the corresponding group numbers are given in the upper panel in Table 4. Based on the ordering, 16 hypotheses/models[4] can be distinguished, see Table 4 under 'Model', where a number represents the group number and a comma separates two subsets. For example, {24135} represents the classical null (H_0), {2, 4, 1, 3, 5} equals the classical alternative (H_A), and {2, 41, 35} denotes $\mu_2,\ \mu_4 = \mu_1,\ \mu_3 = \mu_5$. Table 4 additionally displays the AIC penalty (q_m, the number of distinct group means plus one for the unknown σ^2), the log-likelihood (log L_m), and the PCIC-AIC values (= $-2 \log L_m + 2\, q_m$) for each of the 16 models. The hypothesis with the lowest PCIC-AIC value is the preferred one. This is model 7 with group structure {2, 41, 35} or $\mu_2,\ \mu_4 = \mu_1,\ \mu_3 = \mu_5$.

[4] In model selection, hypotheses are often referred to as models. Therefore, we will use both terms interchangeably in discussing model selection.

Although PCIC-AIC does not yield inconsistencies, it still does not give clear information when examining a set of hypotheses containing at least one order-restricted hypothesis, like H_1 and H_2 in the example. This problem is solved by evaluating the set of theory-based hypotheses directly.

2.3.2. Confirmation

In confirmatory model selection, one evaluates the set containing solely the hypotheses of interest (e.g., those derived from theory such as H_1 and H_2 in our example). One could use the AIC, but often theories lead to order-restricted hypotheses and in that case this AIC is not appropriate. The ORIC (Anraku, 1999), on the other hand, is a modification of the AIC designed to evaluate order-restricted hypotheses as well. We will therefore employ the ORIC rather than the AIC; but note that in the absence of order restrictions, the ORIC reduces to the AIC. Like the AIC, the ORIC comprises a fit term and a penalty term. The fit is also here based on the group means that maximize the likelihood subject to the restrictions in the hypothesis at hand, but now the restrictions may contain inequality constraints ('<' and/or '>'). For instance, let the theory-based hypothesis be H_{C1}: $\mu_1 < \mu_2 < \mu_3$. When the sample means are 1, 2, and 4 and thus in accordance with H_{C1}, the means maximizing the likelihood subject to H_{C1} (referred to as the order-restricted means) are equal to the sample means. When the sample means are not in accordance with H_{C1}, the order-restricted means cannot by definition be equal to the sample means. In that case, the order-restricted means are adjusted samples means such that they not only comply with the restrictions but also maximize the likelihood given those restrictions. When, for example, the sample means are 1, 4, and 2, the sample means are not in accordance with H_{C1} since the second sample mean is not smaller than or equal to the third. Maximizing the likelihood subject to the restrictions in H_{C1} provides order-restricted means that are the (weighted with sample size) averages of these sample means. Hence, assuming equal sample sizes per group, the order-restricted means are 1, $(4 + 2)/2 = 3$, and $(4 + 2)/2 = 3$, respectively. Another difference with respect to AIC is that the penalty, in the presence of order restrictions, equals the expected number of distinct parameters. For example, the expected number of distinct parameters of H_{C2}: $\mu_1 = \mu_2 < \mu_3$ is 2.5 in the case of equal group sizes. Namely, 1 for the unknown σ^2, 1 for the distinct value of $\mu_1 = \mu_2$, and an additional 0.5 for $\mu_2 < \mu_3$. The latter can be deduced from the fact that, under the null where all means are equal, the sample means will be in accordance with $\mu_2 < \mu_3$ half of the time in the case of equal group sizes, leading to two distinct order-restricted mean values; in the other half, the order-restricted means are set equal, leading to one distinct value. That is, the expected number of distinct mean values for H_{C2} is $0.5 \times 2 + 0.5 \times 1 = 1.5$. In practice, the penalty is often hard to determine by hand, but can easily be simulated (Silvapulle & Sen, 2005, pp. 78–81). For instance, the value of the penalty for H_{C1} equals $2\frac{5}{6}$. Notably, this is lower than that for H_A (i.e., $1 + 3 = 4$), since we restrict the parameters, and it is higher than that for H_0 (i.e., $1 + 1 = 2$), since our constraints are not that strict.

In contrast to the $\bar{F}$ test, the ORIC can evaluate multiple theory-based hypotheses simultaneously. It should be stressed that the traditional alternative H_A should be included in the set as a safeguard for weak hypotheses (Kuiper & Hoijtink, 2010), that is, hypotheses not supported by the data. Namely, when all hypotheses are weak, H_A will receive the most support. In the Lucas example, we will evaluate the traditional null H_0 as well, for illustrative purposes, since it is also used in exploration. Hence, the following set of hypotheses is employed:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5, \\
H_1 &: \mu_2 < \{\mu_1, \mu_4\} < \mu_3 = \mu_5, \\
H_2 &: \mu_2 < \mu_5 = \mu_4 < \mu_1 < \mu_3, \\
H_A &: \mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5.
\end{aligned} \tag{3}
$$

Note that, in exploration with $k = 5$, 16 hypotheses are evaluated, whereas the researcher often has a limited number of competing hypotheses under serious consideration; here only four. In Table 5, the values for the penalty term (q_m, the expected number of distinct model parameters), the log-likelihood values ($\log L_m$), and the ORIC values ($= -2 \log L_m + 2q_m$) are given for the four a priori specified hypotheses denoted in equation (3). Note that for hypotheses with no inequalities the ORIC reduces to the AIC, as for models 1 (H_0) and 16 (H_A) in Table 4. In the Lucas example, hypothesis H_1 is preferred, since it has the smallest ORIC value, and thus the best trade-off between fit (likelihood) and complexity (penalty).

2.4. Bayesian model selection

In BMS, the selection of the best hypothesis is not based on an information criterion but on the marginal likelihood of the hypotheses, which is a measure of the degree of support for a hypothesis provided by the data. To interpret several marginal likelihoods at once, it can be helpful to transform them into PMPs. A PMP is the probability that, given the data, the corresponding hypothesis is the best of the set of hypotheses (assuming that all the hypotheses have equal a priori probabilities). The marginal likelihood depends on the likelihood and a prior. A prior reflects pre-existing knowledge or a belief with respect to the parameters (e.g., means). For an elaboration on the role of priors see, for instance, Gelman (2002) and Gelman (2012). In this paper, we use, based on Klugkist, Laudy, and Hoijtink (2005), the normal distribution with a data-based mean and a large variance for every μ_i ($i = 1, \ldots, k$). The prior mean and variance depend not only on the data, but also on a user-specified term (PV) that reflects the vagueness of the prior (see Kuiper & Hoijtink, 2010; Kuiper, Klugkist, & Hoijtink, 2010), where a higher PV value corresponds to an increasing prior vagueness. Klugkist and Hoijtink (2007) show that, if a hypothesis does not contain equality constraints ('='), the relative support of this hypothesis with respect to the unconstrained hypothesis (or other hypotheses without equalities) is not sensitive to the choice of the prior. In contrast, in a set of hypotheses with at least one equality restriction (in at least one hypothesis), the results do depend on the prior

Table 5. Results of model selection (ORIC) and Bayesian model selection (PMP) for the four specified hypotheses

Model H_m	q_m	$\log L_m$	$ORIC_m$	PMP_m
H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	2.00	−292.27	588.54	.00
$\mathbf{H_1}$: $\boldsymbol{\mu_2 < \{\mu_1, \mu_4\} < \mu_3 = \mu_5}$	3.19	−278.05	**562.48**	**.96**
H_2: $\mu_2 < \mu_5 = \mu_4 < \mu_1 < \mu_3$	3.14	−281.76	569.80	.02
H_A: $\mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5$	6.00	−278.05	568.10	.01

Note. ORIC = order-restricted information criterion, with $ORIC_m = -2 \log L_m + 2q_m$; PMP = posterior model probability. Bold type indicates the preferred hypothesis: the lowest ORIC value in model selection and the highest PMP value in Bayesian model selection.

specification. In that case, the vaguer the prior (i.e., a higher *PV* value), the more support for hypotheses with equality constraints. This is known as Lindley's or Bartlett's paradox. Klugkist and Hoijtink (2007) discuss this paradox in the context of BMS in detail. In addition, they show that for reasonable choices of *PV*, the prior sensitivity does usually not lead to a different evaluation of the hypotheses. Although Klugkist *et al.* (2005) designed BMS for the evaluation of order-restricted hypotheses in a confirmatory manner, it can be used in an exploratory fashion as well.

2.4.1. Exploration

One can, for instance, evaluate the 2^{k-1} configurations based on ordered sample means (comparable with the PCIC). The last column in Table 4 shows the results (i.e., PMPs) for the example, when examining this exploratory-based set of hypotheses. In BMS, the hypothesis with the highest PMP value is the preferred one. Hence, model 7 with group structure {2, 41, 35}, that is, μ_2, $\mu_4 = \mu_1$, $\mu_3 = \mu_5$, is the preferred hypothesis in exploration. Note that the same conclusion was obtained before with the PCIC-AIC.

2.4.2. Confirmation

In confirmation, a limited set of well-defined, theory-based hypotheses, such as that in equation (3) above, is evaluated. The results of the confirmatory model selection are revealed in Table 5. The PMP values in the final column indicate that H_1 is the preferred hypothesis (by some margin). Notably, the same conclusion was obtained with the ORIC.

3. Performance of the six techniques

The performance of the six techniques (summarized in Table 1) is evaluated by conducting a simulation study. The performance of hypothesis testing techniques can be measured by statistical power: the probability that the test will reject a false null hypothesis. Thus, statistical power is the ability to detect a true effect. In model selection, one can employ an equivalent of statistical power, namely the probability that the technique will render the most support for the correct or best hypothesis. In the simulation study, the performance is therefore quantified by the number of times the technique prefers the correct or best hypothesis, which is referred to as the true hypothesis rate.

In this section, three comparisons are made: (1) between the performance of hypothesis testing, model selection using information criteria, and BMS; (2) between the performance of methods when evaluating hypotheses with and without order restrictions; and (3) between the performance of methods when evaluating all possible equality-restricted hypotheses versus a solely theory-derived set of hypotheses. Before describing the results, we discuss the chosen values for the number of groups (k) and observations per group (n_i), the hypotheses, and the population parameter values employed in the simulation.

3.1. The number of groups and observations

Techniques able to evaluate order-restricted hypotheses have added value when comparing three or more means, since order-restricted inference with two means reduces to the trivial case of a one-sided test. To obtain a first insight into the performance

Table 6. *Hypotheses tested for $k = 3$ and $k = 5$ in the exploratory ($H_{\text{E}\cdot}$) and confirmatory ($H_{\text{C}\cdot}$) approach*

	Exploration: inspect all possibilities	Confirmation: inspect theory/theories
$k = 3$	H_0: $\mu_1 = \mu_2 = \mu_3$	H_0: $\mu_1 = \mu_2 = \mu_3$
	$H_{\text{E}1}$: $\mu_1 = \mu_2, \mu_3$	$H_{\text{C}1}$: $\mu_1 < \mu_2 < \mu_3$
	$H_{\text{E}2}$: $\mu_1, \mu_2 = \mu_3$	$H_{\text{C}2}$: $\mu_1 = \mu_2 < \mu_3$
	$H_{\text{E}3}$: $\mu_1 = \mu_3, \mu_2$	$H_{\text{C}3}$: $\mu_1 < \mu_2 > \mu_3$
	H_{A}: μ_1, μ_2, μ_3	H_{A}: μ_1, μ_2, μ_3
$k = 5$	H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
	$H_{\text{E}1}$: $\mu_1 = \mu_2 = \mu_3 = \mu_4, \mu_5$	$H_{\text{C}1}$: $\mu_5 = \mu_3 > \{\mu_1, \mu_4\} > \mu_2$
	$H_{\text{E}2}$: $\mu_1 = \mu_2 = \mu_3 = \mu_5, \mu_4$	$H_{\text{C}2}$: $\mu_3 > \mu_1 > \mu_4 = \mu_5 > \mu_2$
	$\vdots$	H_{A}: $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$
	$H_{\text{E}49}$: $\mu_1, \mu_2, \mu_4, \mu_3 = \mu_5$	
	$H_{\text{E}50}$: $\mu_1, \mu_3, \mu_2, \mu_4 = \mu_5$	
	H_{A}: $\mu_1, \mu_2, \mu_3, \mu_4, \mu_5$	

of the six techniques, we start with a simulation with $k = 3$ groups (for both $n = 20$ and $n = 50$)[5] and $k = 5$ groups (for $n = 30$).[6] From these two simulations, a pattern becomes clear with respect to the performance of methods in the exploratory and confirmatory approach and that of those evaluating (reasonable) order-restricted hypotheses or not.

3.2. Hypotheses

Table 6 shows the hypotheses of interest in the simulation study for $k = 3$ and $k = 5$. In the exploratory approach, the hypotheses to be examined are certain group structures represented by pairwise equality ('=') and non-equality ('$\neq$') restrictions. As explained in the previous section, in the exploratory approach (where we use the SWFq test, PCIC-AIC, and BMS), not all possible configurations of means in the observed data set are evaluated, but a subset based on the ordering of the sample means of the data set at hand (e.g., for PCIC-AIC, 4 out of 5 for $k = 3$ and 16 out of 52 for $k = 5$). Nevertheless, in the simulation, more configurations of means can be examined, since the ordering of the sample means may differ by data set in the simulation. Notably, combining the significant and non-significant pairs of means resulting from the SWFq test can lead to favouring one of the hypotheses in the first column of Table 6 or can give inconsistencies. In contrast, PCIC-AIC and BMS always result in preferring one of the hypotheses in Table 6.

In the confirmatory approach (where we here use the $\bar{F}$, ORIC, and BMS), the hypothesis or hypotheses to be tested or selected need to be specified by the researcher. This can be based on previous research or theory. Table 6 displays the hypotheses that are

[5] The results for $n = 20$ are not shown here, since the patterns are the same as for $n = 50$. The only difference is that the performance itself is lower.

[6] Note that the performance of a technique improves when the number of observations per group increases. As a consequence, it is more interesting to examine data sets with small to medium group sizes, that is, group sizes that can detect a large and medium effect, respectively. According to Cohen (1992), with a power of .80 and an α of .05, one needs (at least) 21 observations per group to detect a large effect size in an ANOVA model with three groups, and 52 for a medium effect size. For five groups, one needs (at least) 16 and 39 observations per group, respectively. Based on these figures, we chose to examine $k = 3$ groups for both $n = 20$ and $n = 50$ and (also because of the Lucas example) $k = 5$ groups for $n = 30$.

evaluated in this paper. Bear in mind that the type and number of hypotheses are examples. We choose to include the classical null and alternative and some hypotheses with order restrictions, since this is an advantage of the techniques we examine here. For $k = 3$, we choose to evaluate five hypotheses. This number equals the number of possible configurations of means in exploration, but the structure is different. The set for $k = 5$ is based on Lucas (2003) and consists of the four hypotheses presented in equation (3). Note that, for $k = 5$, 16 hypotheses are evaluated for one data set in exploratory model selection, whereas the researcher is typically only interested in a limited number of (order-restricted) hypotheses.

Hence, the simulation results for $k = 3$ will provide insight into the gain in statistical 'power' when evaluating hypotheses with inequality constraints versus those without. The study with $k = 5$ will additionally investigate the effect of evaluating fewer, theory-derived hypotheses.

The $\bar{F}$ test is designed for testing one theory-based (order-restricted) hypothesis, such as H_{C1}. One can choose to test both H_0 against H_{C1} and H_{C1} against H_A, in addition to H_0 against H_A. The decision rules for these three $\bar{F}$ tests are rather straightforward. However, if M theory-based hypotheses are evaluated by $1 + 2M\,\bar{F}$ tests, the decision rules become very ad hoc and more than one plausible set of decision rules exists. Moreover, no direct comparison is possible between the M theory-based hypotheses. As a consequence, we will only examine the performance of the $\bar{F}$ test for one theory-based (order-restricted) hypothesis, namely H_{C1}.

3.3. Populations

Several populations based on the general ANOVA model are considered. In all populations, the population standard deviation σ is set equal to 1. Sets of population means are given in Table 7. The values are based on the number of groups (k), the true hypothesis, and the effect size given by

$$ES = \frac{1}{\sigma}\sqrt{\frac{1}{k}\sum_{i=1}^{k}(\mu_i - \bar{\mu})^2}, \tag{4}$$

with $\bar{\mu} = \frac{1}{k}\sum_{i=1}^{k}\mu_i$. According to Cohen (1992), an effect size of $ES = 0.10$ is small, of $ES = 0.25$ is medium, and of $ES = 0.40$ is large. Two types of populations can be

Table 7. Population group means (μ_i for $i = 1, \ldots, k$) for zero, small, medium, and large effect size (i.e., $ES = 0.00$, 0.10, 0.25, and 0.40, respectively)

k	True H_m	ES	μ_1	μ_2	μ_3	μ_4	μ_5
3	H_0	0.00	0.000	0.000	0.000		
	H_{C1}	0.10	−0.122	0.000	0.122		
	H_{C1}	0.25	−0.306	0.000	0.306		
	H_{C1}	0.40	−0.490	0.000	0.490		
5	H_0	0.00	0.000	0.000	0.000	0.000	0.000
	H_{C1}	0.10	0.000	−0.122	0.130	0.122	0.130
	H_{C1}	0.25	0.000	−0.306	0.321	0.306	0.321
	H_{C1}	0.40	0.000	−0.490	0.516	0.490	0.516

Table 8. Proportion of times a hypothesis is preferred in exploration for k = 3

k = 3 and n = 50							
ES	Technique	H_0	H_{E1}	H_{E2}	H_{E3}	H_A	Inconsistent
0.00	SWFq	**.944**	.012	.009	.007	.000	.028
0.00	PCIC-AIC	**.653**	.109	.116	.117	.005	–
0.00	BMS-PV2	**.807**	.060	.063	.059	.011	–
0.10	SWFq	.845	.042	.037	.004	**.000**	.072
0.10	PCIC-AIC	.445	.252	.231	.060	**.012**	–
0.10	BMS-PV2	.631	.017	.181	.116	**.055**	–
0.25	SWFq	.227	.268	.262	.000	**.043**	.200
0.25	PCIC-AIC	.040	.335	.337	.002	**.226**	–
0.25	BMS-PV2	.066	.431	.433	.000	**.070**	–
0.40	SWFq	.006	.260	.281	.000	**.416**	.037
0.40	PCIC-AIC	.000	.139	.149	.000	**.712**	–
0.40	BMS-PV2	.001	.256	.279	.000	**.464**	–

Note. SWFq = Shaffer–Welch Fq test; PCIC-AIC = paired-comparison information criterion applied to the Akaike information criterion; BMS-PV2 = Bayesian model selection with PV = 2. Bold face indicates the true hypothesis rate.

distinguished: one where all the population means are identical (ES = 0) and one where they are in accordance with H_{C1} (ES = 0.10, 0.25, 0.40). Based on each population in Table 7, 1,000 data sets are simulated. Subsequently, the corresponding hypotheses in Table 6 are evaluated in each of these data sets. Note that, in exploration when $ES > 0$, H_A and H_{E49} are the correct hypotheses for k = 3 and k = 5, respectively.

3.4. Results

3.4.1. Exploration for k = 3

Table 8 displays the proportions of times the hypotheses are preferred for each technique and each population (i.e., effect size). The SWFq test chooses, as expected, H_0 as the preferred hypothesis about 95% of the time when it is indeed true. Notably, the SWFq test is designed such that $\alpha = .05$. The other two techniques, PCIC-AIC and BMS for PV = 2 (BMS-PV2), do not choose H_0 as often. For ES = 0.10, all three techniques lack power to prefer the correct hypothesis, as is to be expected for a small effect size. When ES = 0.25 or ES = 0.40, the performance is still questionable, but PCIC-AIC clearly outperforms the other two. Thus, the SWFq test performs well under H_0 and PCIC-AIC under H_A, as to be expected owing to their purpose.[7]

[7] We stress that the comparison between the three methods (i.e., hypothesis testing, model selection using information criteria, and BMS) should be done with great care. It is not fair to compare the techniques in terms of power when their Type I error rates are not equal, and *vice versa*. Bear in mind that there is a trade-off between the Type I error rate and power. For example, in Table 8, the true difference in power between the SWFq test and PCIC-AIC can be obtained by inspecting the SWFq test with a nominal α level of approximately .35, since the true hypothesis rate of PCIC-AIC for H_0 under H_0 is about .65. As mentioned before, we merely want to show the (true) hypothesis rates for each of the techniques when employed as commonly done in practice. Again, we then expect the true hypothesis rate of the null hypothesis to be the highest for hypothesis testing and that of non-null hypotheses for model selection, as is the case here.

Table 9. Proportion of times a hypothesis is preferred in confirmation for $k = 3$ (H_0, H_{C1}, and H_A)

$k = 3$ and $n = 50$

ES	Technique	H_0	H_{C1}	H_A	Inconsistent
0.00	$\bar{F}$	**.912**	.052	.033	.003
0.00	ORIC	**.724**	.187	.089	–
0.00	BMS-PV2	**.881**	.075	.044	–
0.10	$\bar{F}$	.682	**.310**	.007	.001
0.10	ORIC	.413	**.563**	.025	–
0.10	BMS-PV2	.615	**.354**	.031	–
0.25	$\bar{F}$	.114	**.886**	.000	.000
0.25	ORIC	.017	**.981**	.002	–
0.25	BMS-PV2	.072	**.917**	.011	–
0.40	$\bar{F}$	.004	**.995**	.001	.000
0.40	ORIC	.000	**.999**	.001	–
0.40	BMS-PV2	.000	**.998**	.002	–

Note. ORIC = order-restricted information criterion; BMS-PV2 = Bayesian model selection with $PV = 2$. Bold face indicates the true hypothesis rate.

3.4.2. Confirmation for k = *3*

In Table 9, the performance of the three techniques suitable for order restrictions is presented for one theory-based, order-restricted hypothesis in addition to the classical null and alternative hypothesis. The $\bar{F}$ test chooses H_0 as the preferred hypothesis about 90% of the time when it is indeed true. This is to be expected, since we perform two tests with respect to H_0 with $\alpha = .05$ and do not correct for multiple testing. The performance of BMS for $PV = 2$ (BMS-PV2) resembles that of the $\bar{F}$ test, while the ORIC performs less well under H_0. For $ES = 0.10$, the three techniques do not perform very well, as is to be expected with a small effect size. For $ES = 0.25$ and $ES = 0.40$, all three techniques perform very well, all preferring H_{C1} more than 88% of the time. For all $ES > 0$, the ORIC performs (somewhat) better than the other two. We refer the reader again to footnote 7 on comparisons of the three methods, which applies to all the comparisons of the three methods throughout this paper.

Table 10 shows the results of evaluating multiple theory-based, order-restricted hypotheses. Since the $\bar{F}$ test is hard to use if more than one theory-based hypothesis is evaluated, that test is excluded. To illustrate the effect of prior vagueness on the results of BMS, we report on the performance of BMS for $PV = 1$ (BMS-PV1), 2 (BMS-PV2), and 3 (BMS-PV3). This table shows that, out of the three, BMS-PV3 most often has the highest support for H_0 and BMS-PV1 the least, as is to be expected with a set of hypotheses containing equality restrictions owing to Lindley's paradox. Additionally, out of the three, the performance of BMS-PV1 resembles that of the ORIC the most. Furthermore, the table shows that, when H_0 is true, BMS-PV3 performs better than the other three techniques and that, when H_{C1} is true, the ORIC outperforms the others. Comparing this table to Table 9, one can see that including more hypotheses decreases the proportion of times the correct hypothesis is chosen.

3.4.3. Evaluating order versus equality restrictions

Comparing Table 10 to Table 8, it is evident that techniques evaluating hypotheses with (reasonable) inequalities have more 'power' than those that only inspect

Table 10. Proportion of times a hypothesis is preferred in confirmation for $k = 3$ (H_0, H_{C1}, H_{C2}, H_{C3}, and H_A)

$k = 3$ and $n = 50$						
ES	Technique	H_0	H_{C1}	H_{C2}	H_{C3}	H_A
0.00	ORIC	**.661**	.070	.116	.115	.038
0.00	BMS-PV1	**.637**	.056	.161	.109	.037
0.00	BMS-PV2	**.777**	.025	.114	.063	.021
0.00	BMS-PV3	**.853**	.014	.078	.040	.015
0.10	ORIC	.353	**.294**	.299	.047	.007
0.10	BMS-PV1	.327	**.248**	.345	.074	.006
0.10	BMS-PV2	.461	**.153**	.333	.052	.001
0.10	BMS-PV3	.537	**.132**	.290	.039	.002
0.25	ORIC	.014	**.744**	.228	.014	.000
0.25	BMS-PV1	.011	**.709**	.254	.026	.000
0.25	BMS-PV2	.039	**.634**	.297	.030	.000
0.25	BMS-PV3	.070	**.593**	.316	.020	.001
0.40	ORIC	.000	**.944**	.056	.001	.000
0.40	BMS-PV1	.000	**.910**	.090	.000	.000
0.40	BMS-PV2	.000	**.888**	.109	.003	.000
0.40	BMS-PV3	.001	**.875**	.122	.002	.000

Note. ORIC = order-restricted information criterion; BMS-PVx = Bayesian model selection with $PV = x$. Bold face indicates the true hypothesis rate.

equalities. For instance, for medium effect sizes (i.e., $ES = 0.25$), the ORIC selects H_{C1} in about 74% of the simulated data sets and the PCIC-AIC selects H_A in about 23% of cases.[8] Bear in mind that H_A is the 'correct' hypothesis in exploration when H_{C1} is the true hypothesis.

3.4.4. Exploration for k = 5

In exploration, when $k = 5$, there are 52 possible hypotheses based on ordered sample means when inspecting more than one data set. Because of the large number of hypotheses, only the results of three of these are given and the results of the other hypotheses are combined (in the 'Other' column in Table 11). We display the results for the null hypothesis H_0, the alternative hypothesis H_A, and the correct hypothesis H_{E49}. Furthermore, we do not include BMS, since evaluating 16 hypotheses per data set with BMS is very time-consuming. Moreover, given the results of the other two techniques, we do not expect that examining BMS would yield additional information.

The proportions of times the hypotheses are selected are displayed in Table 11. This shows that in exploration, if H_0 is true, H_0 is frequently preferred when using the SWFq test (around 95% of the time). In contrast, if H_{E49} is true, the true hypothesis is not chosen by the SWFq test. PCIC-AIC gives the most support to H_0 only 35% of the time when it is true, and to H_{E49} less than 2% of the time when it is true. Hence, exploratory techniques

[8] Note that, in this case, comparison in terms of 'power' is appropriate since both techniques not only serve the same purpose but also have approximately the same Type I error rate.

Table 11. Proportion of times a hypothesis is preferred in exploration for $k = 5$

$k = 5$ and $n = 30$

ES	Technique	H_0	H_{E49}	H_A	Other
0.00	SWFq	**.947**	.000	.000	.053
0.00	PCIC-AIC	**.349**	.000	.000	.651
0.10	SWFq	.878	**.000**	.000	.122
0.10	PCIC-AIC	.201	**.000**	.000	.799
0.25	SWFq	.371	**.000**	.000	.629
0.25	PCIC-AIC	.015	**.002**	.000	.983
0.40	SWFq	.007	**.000**	.000	.993
0.40	PCIC-AIC	.000	**.016**	.000	.984

Note. SWFq = Shaffer–Welch Fq test; PCIC-AIC = paired-comparison information criterion applied to the Akaike information criterion. Bold face indicates the true hypothesis rate.

perform poorly under non-null hypotheses such as H_{E49}. Note that H_A is never the preferred hypothesis.

Table 11 shows that the power to test any specific configuration of means is very low whereas the power to detect at least one effect (1 minus first column) is not. Hence, employing these techniques will usually give one or more statistically significant results but the pattern of significance will vary across data sets. This was also discussed by Maxwell (2004), who distinguished between power to detect any specific comparison (any-pairs power) and power to detect the true pattern of differences (all-pairs power).

3.4.5. *Confirmation for* k = 5

Table 12 shows the results of the three techniques able to examine order-restricted hypotheses for evaluating one theory-based, order-restricted hypothesis. This table exhibits the same patterns as that for $k = 3$, that is, $\bar{F}$ and BMS-PV2 outperform the ORIC under H_0, whereas the ORIC has more power to detect small and medium effect

Table 12. Proportion of times a hypothesis is preferred in confirmation for $k = 5$ (H_0, H_{C1}, and H_A)

$k = 5$ and $n = 30$

ES	Technique	H_0	H_{C1}	H_A	Inconsistent
0.00	$\bar{F}$	**.920**	.045	.031	.004
0.00	ORIC	**.752**	.162	.086	–
0.00	BMS-PV2	**.965**	.027	.008	–
0.10	$\bar{F}$	.715	**.264**	.019	.002
0.10	ORIC	.421	**.536**	.043	–
0.10	BMS-PV2	.779	**.200**	.021	–
0.25	$\bar{F}$	.130	**.860**	.007	.003
0.25	ORIC	.026	**.935**	.039	–
0.25	BMS-PV2	.201	**.768**	.031	–
0.40	$\bar{F}$	.002	**.990**	.008	.000
0.40	ORIC	.000	**.965**	.035	–
0.40	BMS-PV2	.005	**.970**	.025	–

Note. ORIC = order-restricted information criterion; BMS-PV2 = Bayesian model selection with $PV = 2$. Bold face indicates the true hypothesis rate.

Table 13. Proportion of times a hypothesis is preferred in confirmation for $k = 5$ (H_0, H_{C1}, H_{C2}, and H_A)

$k = 5$ and $n = 30$					
ES	Technique	H_0	H_{C1}	H_{C2}	H_A
0.00	ORIC	**.719**	.111	.111	.063
0.00	BMS-PV1	**.874**	.039	.053	.034
0.00	BMS-PV2	**.948**	.024	.018	.010
0.00	BMS-PV3	**.974**	.009	.013	.004
0.10	ORIC	.398	**.365**	.199	.039
0.10	BMS-PV1	.617	**.267**	.087	.029
0.10	BMS-PV2	.761	**.152**	.074	.013
0.10	BMS-PV3	.804	**.140**	.052	.004
0.25	ORIC	.028	**.771**	.164	.036
0.25	BMS-PV1	.083	**.782**	.088	.047
0.25	BMS-PV2	.181	**.690**	.101	.028
0.25	BMS-PV3	.212	**.688**	.076	.024
0.40	ORIC	.000	**.904**	.066	.029
0.40	BMS-PV1	.000	**.913**	.049	.038
0.40	BMS-PV2	.002	**.918**	.035	.045
0.40	BMS-PV3	.005	**.950**	.023	.022

Note. ORIC = order-restricted information criterion; BMS-PVx = Bayesian model selection with $PV = x$. Bold face indicates the true hypothesis rate.

sizes. For large effect sizes, all three techniques perform very well and approximately equally.

Table 13 depicts the performance of the ORIC and BMS (for $PV =$ 1, 2, and 3) for examining two theory-based, order-restricted hypotheses. Tables 12 and 13 show that adding a hypothesis lowers the performance of the techniques. Nevertheless, the trend remains the same, as was the case for $k = 3$. In summary, BMS-PV3 performs better than the other three techniques under H_0 (Lindley's paradox), whereas the ORIC has more power to detect the correct non-null hypothesis (H_{C1}) for a small effect size. In addition, the ORIC and BMS-PV1 outperform the others for a medium effect size; and all techniques perform equally well for a large effect size.

3.4.6. Examining theory-based hypotheses

From Tables 8 and 9, one could see that techniques evaluating hypotheses with (reasonable) inequalities have more 'power' than those solely inspecting equality restrictions. Comparing Table 11 to Table 12 or Table 13, one can see the additional effect of investigating a theory-based set instead of exploring all possible equalities based on ordered sample means. Tables 11 and 12 show that the hypothesis testing techniques both yield high true null hypothesis rates; as they are designed to do. Furthermore, the tables show that the exploratory one (i.e., the SWFq test) gives no support to the true non-null hypothesis, while the confirmatory one (i.e., the $\bar{F}$ test) does support the true non-null hypothesis and this support (logically) increases with effect size. Tables 11 and 13 show that confirmatory model selection (i.e., the ORIC) not only has a (much) higher true null hypothesis rate than its exploratory counterpart (i.e., the PCIC-AIC), but also a much higher true non-null hypothesis rate.

3.5. Discussion

When interest lies in one or more specific hypotheses derived from theory, exploration has some disadvantages. First, the hypotheses of interest are not evaluated directly and often not indirectly either. Moreover, exploratory approaches can generate results that are apparently inconsistent or difficult to interpret. Finally, techniques used in an exploratory fashion exhibit low power to detect specific configurations of means, especially when the number of groups (k) increases.

From the simulations, it can be concluded that techniques employed in a confirmatory way (and able to handle order restrictions) outperform those used in an exploratory way, if interest lies in one or more theory-based (and therefore often order-restricted) hypotheses. There are two reasons for higher rates of true hypothesis selection: (1) specific (often order-restricted) hypotheses are evaluated (analogously to one-sided vs. two-sided testing); and (2) in confirmatory (Bayesian) model selection, a smaller set of hypotheses is examined than in exploratory (Bayesian) model selection.

The $\bar{F}$ test performs very well under H_0. In contrast, it performs less well when another hypothesis is true. Note that one could choose to relax protection of H_0 by testing at a lower α level, which improves the performance when a non-null hypothesis is true. Nevertheless, a disadvantage of the $\bar{F}$ test is that it can evaluate only one theory-based hypothesis (in a straightforward manner). Hence, for a priori theories, we recommend the use of confirmatory model selection, that is, the use of the ORIC or BMS with a limited set of reasonable, theory-based hypotheses. The ORIC performs better than BMS when H_0 is not true (with the exception of large effect sizes, where it is comparable to BMS), whereas BMS is better when H_0 is true (see footnote 7). With respect to prior vagueness, the BMS results with PV values of 1, 2, and 3 show Lindley's paradox as is to be expected with a set of hypotheses including equality constraints. That is, the vaguer the prior, the more support for H_0. Therefore, with $PV = 3$ the false rejection rate of H_0 is controlled best, with $PV = 1$ the power to find the correct (or best) non-null hypothesis is greatest, and with $PV = 2$ a compromise is provided.

4. Robustness of techniques suitable for order restrictions

Not only does confirmation have advantages over exploration, when a researcher has a theory, evaluating reasonable order restrictions can also improve the power to detect the correct (or best) non-null hypothesis. Therefore, more insight should be gained into the performance of techniques able to examine order restrictions (in the case of confirmation). Since little is known about the influence of violations of assumptions of the ANOVA model for these techniques, we will start by investigating their robustness in the presence of heterogeneity.

4.1. Populations and hypotheses

As a starting point, we consider only $k = 3$ groups and the first mentioned order-restricted hypothesis in Table 6 (i.e., H_{C1}). This produces the following set of hypotheses:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \mu_3, \\ &H_{C1} : \mu_1 < \mu_2 < \mu_3, \\ &H_A : \mu_1,\ \mu_2,\ \mu_3. \end{aligned} \tag{5}$$

As in the previous simulation study, H_0 is included for illustrative purposes; often it is not a plausible hypothesis and should not then be included in the set.

According to Schumacker and Akers (2001), Box (1954), and Tabachnick and Fidell (2001), the ANOVA F test is not as robust to heterogeneity when the groups with the largest sample sizes have the highest variances and those with the smallest sample sizes have the lowest variances than when the group sizes are equal. Therefore, we will examine both equal group sizes ($n_1 = n_2 = n_3 = n$, with $n = 20$ and $n = 50$) and unequal group sizes ($n_1 = 20$, $n_2 = 50$, and $n_3 = 100$).

Here, the populations differ from the previous ones, because of the group-specific variances, σ_i^2 for $i = 1, 2, 3$. Although the population standard deviations are divergent, they are assumed to be equal in the ANOVA model. Due to heterogeneity, the effect size is calculated by equation (4), where σ is replaced by the pooled standard deviation

$$\sigma_p = \sqrt{\frac{\sum N_i \sigma_i^2}{\sum N_i}},$$

with N_i the size of group i in the population. In the simulation, we will use n_i instead of N_i, because we assume that the relative sizes of the groups in the samples equal those in the population. By setting $\sigma_p = 1$, the same effect sizes and population means as in the previous simulation study (upper panel in Table 7) are obtained. Again, we have one set where all the population means are identical ($ES = 0$) and H_0 is true and three where they exhibit an upward trend ($ES > 0$) and H_{C1} is true.

To manipulate the severity of (population) heterogeneity of variance in the simulation study, we employ a measure based on Hartley's (1950) heterogeneity test. It is called $F_{\max}$ and equals the ratio of the largest and smallest group variance, that is,

$$F_{\max} = \frac{\sigma^2_{\max}}{\sigma^2_{\min}}. \tag{6}$$

Now, the standard deviations (σ_i) are with $\sigma_p = 1$ solely based on the $F_{\max}$ value in equation (6). An $F_{\max}$ value of one implies that there is no difference in the group variances, that is, the homogeneity of variance assumption is not violated, as was the case in the previous simulation study. Consequently, we will use $F_{\max} = 1$ as the baseline. Evidently, a higher $F_{\max}$ value indicates a larger difference in group variances and *vice versa*. Tabachnick and Fidell (2001) conclude that an $F_{\max}$ value of 10 is acceptable for analyses with equal group sizes and $F_{\max} = 3$ for unequal group sizes. However, Box (1954) shows for $F_{\max} = 3$ that the F test is severely affected when the group sizes and the group variances exhibit opposite trends. Based on these findings, we additionally set $F_{\max}$ to 3 and 10. To examine the effect of a large violation, we examine $F_{\max} = 100$ as well. For $F_{\max}$>1, different orderings of the σ_i values in relation to the ordering of group sizes exist. Note that the ranking of the σ_i values is arbitrary in the case of equal group sizes. Hence, for equal group sizes, we only examine σ_i values with an upward trend, that is, the σ_i^2 values increase with i. If the group sizes are unequal, we will investigate samples with

Table 14. Population standard deviations

		Equal group sizes*			Unequal group sizes**		
Type of trend	$F_{\max}$	σ_1	σ_2	σ_3	σ_1	σ_2	σ_3
Baseline	1	1.000	1.000	1.000	1.000	1.000	1.000
Upward	3	0.707	1.000	1.225	0.612	1.000	1.061
Downward	3				1.500	1.000	0.866
Upward	10	0.426	1.000	1.348	0.343	1.000	1.085
Downward	10				2.000	1.000	0.632
Upward	100	0.141	1.000	1.407	0.109	1.000	1.094
Downward	100				2.390	1.000	0.239

*$n_1 = n_2 = n_3 = n$, with $n = 20$ and $n = 50$. **$n_1 = 20, \quad n_2 = 50, \quad n_3 = 100$.

sizes increasing with i with two rankings of σ_i values based on the results of Box (1954), namely one with an upward trend and one with a downward trend. Owing to the two types of group sizes (equal and unequal), the four $F_{\max}$ values, and the two types of trends (upward or downward), 11 types of sets of population standard deviations are investigated, which are given in Table 14.

For each combination of n_i in relation to σ_i, $F_{\max}$, and ES, 1,000 data sets are simulated. Subsequently, the hypotheses in equation (5) are evaluated in each of these data sets.

4.2. Results and discussion

Figure 1 displays the proportion of times H_{C1} is preferred by the three techniques suitable for examining order-restricted hypotheses ($\bar{F}$ at the top, ORIC in the middle, and BMS at the bottom) for effect size ES (represented by the different lines in each plot) and heterogeneity level $F_{\max}$ (depicted on the x-axis of each plot) for unequal group sizes. The results for equal group sizes are not plotted, since they are very robust, but we will briefly elaborate on these later on. Performance is measured by the proportion of times the correct hypothesis, that is, H_{C1}, is preferred (displayed on the y-axis of each plot). Notably, complete robustness to heterogeneity would imply only horizontal lines. The figure shows that the effect of heterogeneity on the performance of the three techniques, when H_{C1} is true, depends on the effect size. For medium to large effect sizes (i.e., $ES = 0.25$ and .40), the proportion of times H_{C1} is preferred increases with $F_{\max}$, when the σs show an upward trend (see the two top lines in the panels on the left-hand side in Figure 1). In contrast, the proportion decreases with $F_{\max}$, when the σs show a downward trend (see the two top lines the panels on the right-hand side in Figure 1). The opposite holds true for small effect sizes. That is, for $ES = 0$ and 0.10, the proportion of times H_{C1} is preferred decreases (increases) with $F_{\max}$, when the σs show an upward (downward) trend (see the two bottom lines in the panels on the left-hand (right-hand) side in Figure 1). Furthermore, the difference in performance owing to heterogeneity is larger when the σs have a downward trend than when they exhibit an upward trend. It should be stressed that the difference in performance due to heterogeneity is not large for $F_{\max} = 3$ in all cases or for $F_{\max} = 10$ when the σs exhibit an upward trend. Moreover, an $F_{\max}$ value of 100 can be considered an extreme violation, compared to the benchmarks of Tabachnick and Fidell (2001) discussed earlier, where an $F_{\max}$ value of 10 and 3 is concluded to be acceptable for analyses with equal group sizes and unequal group sizes, respectively.

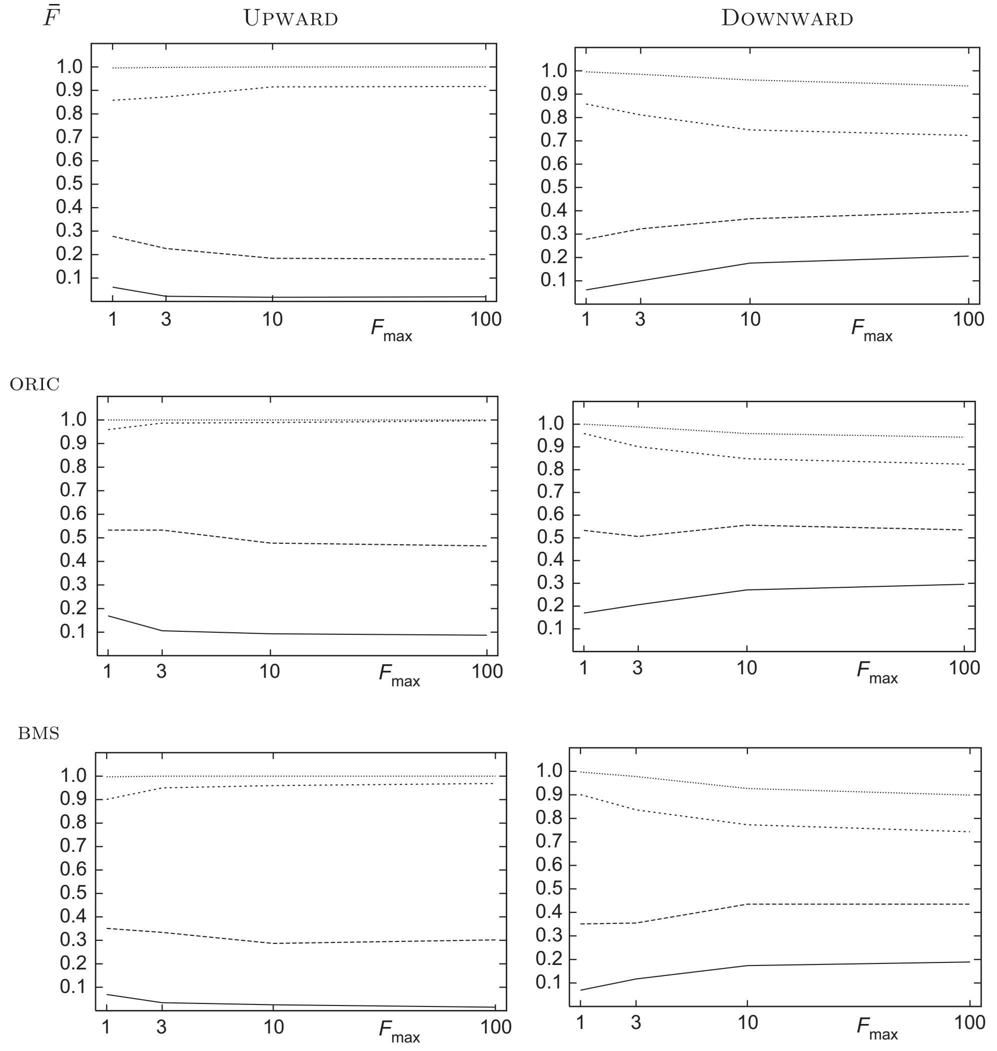

Figure 1. Proportion of times H_{C1} is selected as best hypothesis, by each of the techniques (i.e., $\bar{F}$, ORIC, or BMS), for increasing heterogeneity (F_{max}) for unbalanced designs with either larger variances in larger groups (Upward) or smaller variances in larger groups (Downward). In each plot, the four lines from bottom to top show results for ES = 0, 0.10, 0.25, and 0.40, respectively.

Although the general trend is clear from Figure 1, the magnitude of the deviations from the baseline is less clear. In addition, it does not reveal performance when the group sizes are equal or the proportion of times H_0 is selected. In this simulation, interest lies in robustness of the techniques and not in the performance itself. Therefore, we include two tables in the appendix which present the difference in performance for an $F_{max} > 1$ compared with $F_{max} = 1$. These differences give an indication of the robustness to heterogeneity on the performance under both H_0 and H_{C1}, where a difference of zero indicates full robustness and a higher absolute difference reflects a poorer robustness.

Table A1 shows the difference in performance with respect to H_0 for $ES = 0$. With equal group sizes, no obvious trend can be seen across techniques or $F_{\max}$ values. Additionally, the differences in proportion of times H_0 is preferred are extremely small for the different $F_{\max}$ values. In the unequal sample size condition, the proportion of times H_0 is preferred increases (decreases) with $F_{\max}$ when the σs exhibit an upward (downward) trend. Furthermore, the effects of heterogeneity on performance are more severe for the downward trend than for the upward trend. It should be stressed that these results resemble the effect of heterogeneity on the ANOVA *F* test.

Table A2 shows, for all effect sizes, the differences in the proportion of times H_{C1} is preferred (for $F_{\max} > 1$ compared with $F_{\max} = 1$). These differences do not vary by more than an absolute value of 0.05 for the three $F_{\max}$ values with any effect size when sample sizes are equal for both $n = 20$ and 50. Also no pronounced trend is apparent here. The patterns in robustness for unequal group sizes have already been discussed in relation to Figure 1. In addition, Table A2 shows that the absolute difference in performance is less than 0.10 in case of the upward trend and 0.16 in case of the downward trend.

We conclude that for all three techniques (i.e., the $\bar{F}$ test, ORIC, and BMS) the performance under both H_0 and H_{C1} is robust to heterogeneity when the group sizes are equal. For unequal group sizes and when the group standard deviations exhibit the same trend, the performance under both H_0 and H_{C1} is still quite robust. However, when the group standard deviations exhibit a trend opposite to that of the group sizes, there are larger deviations, especially for $F_{\max}$ values larger than 3. Although more research is required in order to state that these techniques are robust to heterogeneity, our results should encourage their use.

5. Conclusion

In this paper, we have examined the performance and robustness of various statistical methods for the evaluation of hypotheses in the context of analysis of variance models. The focus was on the researcher having theories, in which case hypotheses often contain order restrictions on the means. The main goal of the first set of simulation studies was the comparison of the exploratory versus the confirmatory approach (and the additional effect of being able to evaluate order restrictions) within the three methods: hypothesis testing, model selection using information criteria, and BMS. Results showed that the confirmatory approach outperforms the exploratory approach. Confirmation yields higher true hypothesis rates, since (1) a specific (often order-restricted) hypothesis is evaluated and (2) the set of hypotheses is usually (much) smaller than that in exploration (especially when the number of groups increases). Moreover, the exploratory approach exhibits low power to detect specific configurations of means. Additionally, techniques used in an exploratory manner fail to evaluate the hypothesis of interest directly and often indirectly. When comparing the three methods, the results show (as to be expected) that hypothesis testing techniques have high true null hypothesis rates and (Bayesian) model selection techniques high(er) true non-null hypothesis rates (see footnote 7). Furthermore, hypothesis testing techniques can generate results that are inconsistent or hard to interpret.

The second set of simulation studies addressed the robustness of three techniques suitable for order restrictions (employed in a confirmatory way) in case of heterogeneity. Results showed that all three techniques (i.e., the $\bar{F}$ test, ORIC, and BMS) are robust to heterogeneity when the group sizes are equal and quite robust when the group standard deviations exhibit the same trend as the group sizes. If the group standard deviations exhibit a trend opposite to that of the group sizes, there are larger deviations, especially for $F_{\max}$ values larger than 3.

In this paper, we have investigated six techniques (distinguished by the three methods and being suitable for evaluating order restrictions or not) within the context of a simple ANOVA model. Nevertheless, most of the techniques presented in this paper that are suitable for inspecting order restrictions are also available for extensions of the model (e.g., for multivariate designs). Theoretically, the $\bar{F}$ test can be generalized to test hypotheses in multivariate normal linear models, but to the best of our knowledge this has not yet been done. The ORIC is generalized to a criterion called the GORIC, which can evaluate hypotheses in multivariate normal linear models (Kuiper, Hoijtink & Silvapulle, 2011, 2012). For BMS, extensions to a broad range of multivariate normal models have also been formulated (Mulder, Hoijtink, & Klugkist, 2010). In the multivariate context, it would be interesting to examine the performance and robustness of the various techniques.

To apply the approaches investigated in this paper, tailor-made software is available;[9] both the software and a tutorial are presented in Kuiper and Hoijtink (2010) and Kuiper *et al.* (2010). The code is written in Fortran 90, providing stand-alone and free-to-use software. This is also the case for the BIEMS program (Mulder, Hoijtink & de Leeuw, 2012) enabling BMS for a broad range of multivariate normal linear models with order constraints on the model parameters.[10] Alternatively, GORIC is available via stand-alone and free-to-use software[11] (Kuiper & Hoijtink, 2013) and in the R package `goric` (Gerhard & Kuiper, 2012). Other relevant software includes the R package `contrast` (Kuhn, with contributions from Weston, Wing, & Forester, 2011), which can be used to evaluate a priori specified patterns by assigning weights to each group mean, and the R library `isoregbf`,[12] which can test various order restrictions and is based on the computation of Bayes factors in an ANOVA-like set-up. In Wetzels, Grasman, and Wagenmakers (2012) and Rouder, Morey, Speckman, and Province (2012), the reader can find other work on BMS in an ANOVA setting. Both papers supply R code and the latter have also written an R package called `BayesFactor` (Morey & Rouder, 2013). Note, however, that these two papers do not address order-restricted hypotheses.

To conclude, there are several options for researchers with specific theories that may contain order restrictions. For more than one theory (i.e., competing theories), the two model selection approaches are recommended. Both the ORIC and the BMS approach perform well. They have acceptable error levels, good power to find the best model, and are rather robust to violations of the homogeneity assumption. Researchers with just one order-constrained hypothesis can also choose to conduct the $\bar{F}$ test. Finally, we would emphasize that we do not object to the use of exploratory data analysis; but, if a researcher has specific theories, it is very worthwhile to use a confirmatory approach (with a technique suitable for order restrictions).

[9] http://staff.fss.uu.nl/RMKuiper.

[10] http://vkc.library.uu. nl/vkc/ms/research/ProjectsWiki/Software.aspx.

[11] http://staff.fss.uu.nl/RMKuiper.

[12] https://sites.google.com/site/rosselldavid/software.

References

Anraku, K. (1999). An information criterion for parameters under a simple order restriction. *Biometrika*, *86*, 141–152. doi:10.1093/biomet/86.1.141

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in the one-way classification. *Annals Mathematical Statistics*, *25*, 290–302.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer-Verlag.

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

Cribbie, R. A., & Keselman, H. J. (2000). A power comparison of pairwise multiple comparison procedures: A model comparison approach versus stepwise procedures. Paper presented at the 2000 Annual Meeting of the American Educational Research Association, New Orleans, LO.

Dayton, C. M. (1998). Information criteria for paired-comparison problem. *American Statistician*, *52*, 144–151. doi:10.4135/9781412984720

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods*, *8*, 61–71. doi:10.1037/1082-989X.8.1.61

Gelman, A. (2002). Prior distribution. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics* (Vol. 3, pp. 1634–1637). Chichester, UK: John Wiley.

Gelman, A. (2012, February). *Statistical modeling, causal inference, and social science: What is a prior distribution?* Retrieved from http://andrewgelman.com/2012/02/05/what-is-a-prior-distribution/

Gerhard, D., & Kuiper, R. M. (2012). *goric: Generalized order-restricted information criterion [Computer software manual]*. Retrieved from http://CRAN.R-project.org/package=goric (R package version 0.0-6).

Hartley, H. O. (1950). The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, *37*, 308–312. doi:10.1093/biomet/37.3-4.308

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*, 6367–6379. doi:10.1016/j.csda.2007.01.024

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

Kuhn, M., with contributions from Weston, S., Wing, J., & Forester, J. (2011). *contrast: A collection of contrast methods [Computer software manual]*. Retrieved from http://CRAN.R-project.org/package=contrast (R package version 0.17).

Kuiper, R. M., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, *15*, 69–86. doi:10.1037/a0018720

Kuiper, R. M., & Hoijtink, H. (2013). A Fortran 90 program for the generalized order-restricted information criterion. *Journal of Statistical Software*, *54*, 1–19.

Kuiper, R. M., Klugkist, I., & Hoijtink, H. (2010). A Fortran 90 program for confirmatory analysis of variance. *Journal of Statistical Software*, *34*(8), 1–31.

Kuiper, R. M., Hoijtink, H., & Silvapulle, M. J. (2011). An Akaike-type information criterion for model selection under inequality constraints. *Biometrika*, *98*, 495–501. doi:10.1093/biomet/asr002

Kuiper, R. M., Hoijtink, H., & Silvapulle, M. J. (2012). Generalization of the order-restricted information criterion for multivariate normal linear models. *Journal of Statistical Planning and Inference*, *142*, 2454–2463. doi:10.1016/j.jspi.2012.03.007

Lucas, J. W. (2003). Status processes and the institutionalization of women as leaders. *American Sociological Review*, *68*, 464–480. doi:10.2307/1519733

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*, 147–163. doi:10.1037/1082-989X.9.2.147

Morey, R. D., & Rouder, J. N. (2013). *BayesFactor: Computation of Bayes factors for common designs [Computer software manual]*. Retrieved from http://CRAN.R-project.org/package=BayesFactor (R package version 0.9.2).

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*, 887–906. doi:10.1016/j.jspi.2009.09.022

Mulder, J., Hoijtink, H., & de Leeuw, C. (2012). Biems: A Fortran 90 program for calculating Bayes factors for inequality and equality constrained models. *Journal of Statistical Software*, *46*(2), 1–39.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods*, *7*, 504–523. doi:10.1037/1082-989X.7.4.504

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356–374. doi:10.1016/j.jmp.2012.08.001

Schumacker, R. E., & Akers, A. (2001). *Understanding statistical concepts using S-Plus*. Mahwah, NJ: Erlbaum.

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Inequality, order, and shape restrictions*. Hoboken, NJ: Wiley.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn & Bacon.

Toothaker, L. E. (1993). *Multiple comparison procedures*. Newbury Park, CA: Sage.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078

Wesel, F., Hoijtink, H., & Klugkist, I. (2011). Choosing priors for constrained analysis of variance: Methods based on training data. *Scandinavian Journal of Statistics*, *38*, 666–690. doi:10.1111/j.1467-9469.2010.00719.x

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2012). A default bayesian hypothesis test for anova designs. *American Statistician*, *66*, 104–111.

Received 31 October 2011; revised version received 14 February 2014

Appendix: Results on robustness to heterogeneity of techniques suitable for order restrictions

Table A1. Difference in proportion of times H_0 is selected for $F_{\max}$ = 3, 10, and 100 compared with $F_{\max}$ = 1 for ES = 0

Technique	$n_i = 20$			$n_i = 50$			$n_1 = 20,\ n_2 = 50,\ n_3 = 100$					
							Upward trend			Downward trend		
	3	10	100	3	10	100	3	10	100	3	10	100
$\bar{F}$	−.012	−.017	−.006	.005	−.004	−.051	.043	.057	.054	−.086	−.213	−.293
ORIC	−.005	−.016	.006	.040	.039	−.002	.085	.122	.125	−.114	−.210	−.291
BMS	−.016	−.017	−.013	.011	.009	−.048	.039	.056	.059	−.096	−.207	−.286

Note. *ES* = effect size; ORIC = order-restricted information criterion; BMS = Bayesian model selection.

Table A2. Difference in proportion of times H_{C1} is selected for $F_{\max}$ = 3, 10, and 100 compared with $F_{\max}$ = 1

Technique		$n_i = 20$			$n_i = 50$			$n_1 = 20, n_2 = 50, n_3 = 100$					
								Upward trend			Downward trend		
	ES	3	10	100	3	10	100	3	10	100	3	10	100
$\bar{F}$	0	−.002	.007	−.003	.001	.006	.042	−.039	−.043	−.041	.039	.115	.145
	0.10	.031	.012	.026	−.017	.008	.008	−.052	−.094	−.097	.045	.088	.118
	0.25	.039	.025	−.020	.014	.013	.026	.014	.057	.059	−.047	−.111	−.135
	0.40	−.009	−.002	−.002	.005	.004	.005	.002	.004	.004	−.011	−.035	−.061
ORIC	0	−.015	.004	−.010	−.028	−.037	.012	−.063	−.076	−.082	.037	.102	.127
	0.10	.045	.017	−.017	−.004	.002	−.013	.000	−.055	−.067	−.027	.023	.002
	0.25	.012	.009	−.020	.006	−.001	−.003	.028	.030	.038	−.058	−.111	−.135
	0.40	−.009	.003	−.041	.000	.000	.001	.000	.000	.000	−.012	−.041	−.057
BMS	0	−.003	.006	.001	−.011	−.017	.043	−.035	−.044	−.054	.048	.105	.120
	0.10	.030	.012	.011	.012	.021	.021	−.017	−.064	−.049	.004	.084	.084
	0.25	.038	.038	−.004	.025	.008	.019	.049	.059	.068	−.065	−.128	−.158
	0.40	−.014	−.015	−.007	.000	.000	.002	.003	.003	.003	−.019	−.070	−.098

Note. *ES* = effect size; ORIC = order-restricted information criterion; BMS = Bayesian model selection.