Open Journal of Statistics, 2014, 4, 710-725
Published Online October 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.49066

Confirmatory Methods, or Huge Samples, Are Required to Obtain Power for the Evaluation of Theories

Irene Klugkist[1*], Laura Post[1], Freek Haarhuis[1], Floryt van Wesel[2,3]

[1]Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands
[2]Department of Educational Neuroscience, Faculty of Psychology and Education, VU University Amsterdam, Amsterdam, The Netherlands
[3]Department of Methodology, Faculty of Psychology and Education, VU University Amsterdam, Amsterdam, The Netherlands
Email: *i.klugkist@uu.nl

Received 4 August 2014; revised 8 September 2014; accepted 16 September 2014

Abstract

Experimental studies are usually designed with specific expectations about the results in mind. However, most researchers apply some form of omnibus test to test for *any differences*, with follow up tests like pairwise comparisons or simple effects analyses for further investigation of the effects. The power to find full support for the theory with such an exploratory approach which is usually based on multiple testing is, however, rather disappointing. With the simulations in this paper we showed that many of the common choices in hypothesis testing led to a severely underpowered form of theory evaluation. Furthermore, some less commonly used approaches were presented and a comparison of results in terms of power to find support for the theory was made. We concluded that confirmatory methods are required in the context of theory evaluation and that the scientific literature would benefit from a clearer distinction between confirmatory and exploratory findings. Also, we emphasis the importance of reporting all tests, significant or not, including the appropriate sample statistics like means and standard deviations. Another recommendation is related to the fact that researchers, when they discuss the conclusions of their own study, seem to underestimate the role of sampling variability. The execution of more replication studies in combination with proper reporting of all results provides insight in between study variability and the amount of chance findings.

Keywords

Confirmatory Research, Exploratory Research, Power, Theory Evaluation

*Corresponding author.

How to cite this paper: Klugkist, I., Post, L., Haarhuis, F. and van Wesel, F. (2014) Confirmatory Methods, or Huge Samples, Are Required to Obtain Power for the Evaluation of Theories. *Open Journal of Statistics*, **4**, 710-725.
http://dx.doi.org/10.4236/ojs.2014.49066

1. Introduction

Experimental studies are usually designed with specific expectations about the results in mind. Van den Hout and colleagues, for instance, designed a study to investigate differences in performance between interventions for posttraumatic stress disorder [1]. While it has been shown that Eye Movement Desensitization and Reprocessing (EMDR) is an effective treatment, recently, therapists sometimes replace eye movements (EMs) by alternating beep tones. To investigate if the intervention based on beep tones was: 1) effective at all, and 2) equally effective as the intervention using EMs, patients were randomized over three groups: recall only, recall with EMs, or recall with beep tones. Three competing expectations for the outcome were formulated:

H_1: beep tones are as effective as EMs.

H_2: beep tones are not effective at all.

H_3: beep tones are effective, but not as effective as EMs.

In terms of the three conditions, this can also be expressed as:

H_1: $\{\text{EMs} = \text{beep tones}\} > \text{recall only}$.

H_2: $\text{EMs} > \{\text{beep tones} = \text{recall only}\}$.

H_3: $\text{EMs} > \text{beep tones} > \text{recall only}$.

The main goal of this experiment was to evaluate for which of these three competing hypotheses the data provided most support.

Another illustration of research with specific expectations about the results is presented by [2]. In a study on the effect of stereotype threats on the math performance of women and men, they hypothesized that on a relative simple math test there would be no differences in performance between men and women, but on a difficult test where they expected both men and women to perform worse than on the simple test, they did also expect men to score better than women. Let μ denote the mean performance and the subscripts w = women, m = men, s = simple and d = difficult. The expectations can be expressed as: $H_{\text{Spencer}} : \mu_{w,d} < \mu_{m,d} < \{\mu_{w,s} = \mu_{m,s}\}$.

This is an example of a factorial design but the hypothesis of interest is not formulated as, nor approached by, (default) testing for main or interaction effects, but instead expresses the specific theory of the researcher in one hypothesis.

Both examples show that research expectations are often expressed using order constraints on the model parameters (e.g. means in experimental groups). Hypotheses in terms of such constraints are denoted ordered, inequality constrained, or informative hypotheses [3]. We prefer the last term for two reasons. First, the hypothesis of interest can include order/inequality constraints (<, >), but also equality constraints (=) and unconstrained parts (denoted using a comma, e.g. $\{\mu_1, \mu_2\} > \mu_3$ states that both μ_1 and μ_2 are greater than μ_3, but there is no constraint with respect to the mutual relation of μ_1 and μ_2). Second, it emphasizes that the hypothesis is informative in the sense that it captures the information the researcher is interested in (*i.e.*, the theory or explicit expectation).

A review of empirical literature shows that many research articles contain such hypotheses, that is, in the introduction of the paper the authors clearly state what their expectations with respect to (part of) the outcomes are. This is especially the case in experimental studies. Despite such prespecified expectations or theories, most researchers apply some form of omnibus test to test for *any differences*, with follow up tests like pairwise comparisons or simple effects analyses for further investigation of the effects. The power to find *full support* for the theory with such an approach is, however, rather disappointing.

To illustrate this consider the hypothesis expressing the expectation that four means are of increasing magnitude, that is, the hypothesis states what is called a simple ordering of four means: $H_{\text{informative}} : \mu_1 < \mu_2 < \mu_3 < \mu_4$.

After assuring that all assumptions to perform an analysis of variance (ANOVA) are met, we believe that the majority of researchers would approach this hypothesis by first testing the omnibus F-test to see if there is evidence for any differences between the four means. After rejection of the null hypothesis "all means equal", one would probably investigate the pairwise comparisons to determine which means differ from each other. Throughout the paper, $\alpha = 0.05$ will be used to determine statistical significance.

Full support for the theory could be claimed if 1) the omnibus F-test is statistically significant, 2) the sample means are in the hypothesized order, and 3) the pairwise comparisons testing $H_{01} : \mu_1 = \mu_2$, $H_{02} : \mu_2 = \mu_3$, and $H_{03} : \mu_3 = \mu_4$ are statistically significant. Note that other approaches to decide on full support for $H_{\text{informative}}$ are available and several will be discussed and investigated in the next section.

In a small simulation study, with population means in the hypothesized order, an effect size that can be la-

beled as medium (Cohen's $f = 0.27$) and a sample size of 50 per group, the power of the omnibus ANOVA was 0.92. Stated differently, in 92% of the data sets that were all simulated from the specified population, the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ was rejected ($p < 0.05$). However, the percentage of data sets in which also the sample means were in the hypothesized order (*i.e.*, $M_1 < M_2 < M_3 < M_4$) was only 65%. Worse, when counting the data sets in which also all three pairwise tests were statistically significant ($p < 0.05$), we ended up with a disappointing result of *zero*! The power to find full support for the hypothesized order, where full support is defined as finding statistically significant pairwise differences between the four subsequent means, is 0.00.

From these numbers it is clear that the power for an omnibus ANOVA and the power to find support for a specific expectation about a pattern of means can deviate substantially. Similar results and conclusions were previously reported in [4], although not in the context of testing informative (order constrained) hypotheses but in the general context of multiple testing. In [4] it was argued that multiple testing causes studies to be underpowered and that this leads to inconsistencies in the published literature. Multiple testing is also the main explanation for the low power in our illustration.

This paper has two main goals. We will show that many of the common choices in hypothesis testing lead to a severely underpowered form of theory evaluation. Furthermore, we will compare the results with available but less commonly used approaches and discuss when each of them could serve as a valuable and more powerful alternative.

In the next section six approaches are described that can be used in the context of a one-way ANOVA when the hypothesis of interest is a simple ordering of *k* means. For a variety of populations, two questions are investigated: "What is the power to find full support given that the power for the omnibus test is 80%?", and "What is the required sample size to obtain 80% full support power for the specific expectations?". In Section 3, the results are reported of simulation studies meant to investigate the specific interaction hypothesis in the two-way design of the math performance example just introduced. The paper is concluded with a discussion of results and possible implications for psychological research.

2. One-Way Analysis of Variance

In the context of a one-way ANOVA with *k* groups, six approaches are presented that researchers could employ when evaluating the explicit research hypothesis that the means are increasing, that is: $H_{\text{informative}}$: $\mu_1 < \mu_2 < \cdots < \mu_{k-1} < \mu_k$. Although we do not claim that these are the only options available, we do believe that many researchers will recognize one or more of the presented approaches and probably have employed them in their own research. With simulation studies the performance of the six approaches will be systematically evaluated for a variety of populations. The question that is investigated is: If the hypothesized ordering of means is indeed present in the population, how often will each of the approaches find *full support* for this hypothesis? The technical aspects of approaches I-V (all using NHT) are provided in Appendix A. A short summary of approach VI (a Bayesian approach) is presented in Appendix B.

In Section 2.1, we present three approaches that are frequently seen in published research papers. However, these methods are not the best choice for theory evaluation, that is, for testing explicit hypotheses. Therefore, in Section 2.2, three alternative approaches are presented that may be better suited to evaluate pre-specified explicit hypotheses, but that are probably less familiar to some researchers. Sections 2.3, 2.4, and 2.5 present the results of several simulation studies.

2.1. Three Omnibus Test Based Approaches

The first three approaches are based on performing an omnibus ANOVA, despite the fact that the hypothesis of interest is more specific than the hypothesis evaluated with the omnibus test:

H_0: $\mu_1 = \cdots = \mu_k$.

H_A: not H_0.

Additionally, to evaluate the actual research hypothesis (*i.e.*, to be able to claim full support), three different follow-up procedures are considered.

I. *Omnibus ANOVA + sample means in hypothesized order*

To claim support for the research hypothesis, the omnibus test must be statistically significant ($p < 0.05$) and the sample means (M_j) must be in the hypothesized order. To have convincing evidence for the specific hypothesis and to get the work published it seems, however, necessary to include follow-up testing. This is not yet

done in approach I.

II. *Omnibus ANOVA + sample means in hypothesized order + all pairwise tests for subsequent means significantly different* (*with* $\alpha = 0.5$, *no multiple testing correction*)

If the omnibus test is significant ($p < 0.05$), researchers often continue with pairwise comparisons to further investigate which means differ significantly from each other. Full support for the hypothesis can only be claimed if all subsequent pairs of means (*i.e.*, M_1 with M_2, M_2 with M_3, etc.) are significantly different and in the correct direction ($M_1 < M_2$, $M_2 < M_3$, etc.).

Since we are now applying multiple tests (each with α level 0.05) to get an answer to one specific research question, the issue of inflated type 1 errors emerges. Researchers have to make a decision about how to control the family wise error and make a choice between a long list of available correction methods. This is not yet done in approach II, where no α correction is made. Note that this is equal to using the LSD (Least Significant Difference) method that SPSS offers (see, for instance, [5]).

III. *Omnibus ANOVA + sample means in hypothesized order + all pairwise tests for subsequent means significantly different* (*with Bonferonni corrected* α)

In the third approach, the Bonferonni α correction is applied for the pairwise tests. The Bonferonni correction divides the desired overall α level by the total number of pairwise comparisons. Approaches II and III, therefore, provide results for two extremes: LSD is very liberal (no correction), Bonferonni is rather conservative (stringent correction). Note that default SPSS Bonferonni output is based on the total number of possible tests, that is, $1/2k(k-1)$. However, since we investigate a simple ordering, we only need $(k-1)$ pairwise comparisons and will therefore use a less stringent correction (retaining more power).

2.2. Three Alternative Approaches

To do justice to the confirmatory nature of research, for $H_{\text{informative}}$ an approach that tests the hypothesis more directly would be a better choice. Here, we present three approaches that can be used to evaluate an informative hypothesis that states a simple order of means.

IV. *Multiple planned contrasts* (*one-sided*)

Planned contrast testing is an alternative to omnibus testing and can be used whenever pre-specified hypotheses are available (e.g., [6]). In case of a simple order of k means, one option is to test $k-1$ contrasts, where each contrast C_i $(i = 1, \cdots, k-1)$ represents the pairwise comparison of two subsequent means. The set of contrasts for $k = 6$ means, for instance, is:

C_1:	−1	+1	0	0	0	0
C_2:	0	−1	+1	0	0	0
C_3:	0	0	−1	+1	0	0
C_4:	0	0	0	−1	+1	0
C_5:	0	0	0	0	−1	+1

This provides, for example, $C_1 = -1*M_1 + 1*M_2 + 0*M_3 + 0*M_4 + 0*M_5 + 0*M_6 = M_2 - M_1$. For each contrast, $H_0 : C_i = 0$ is tested against $H_A : C_i > 0$ (*i.e.*, with one sided p-values). With planned contrast testing it is not necessary to first evaluate the omnibus ANOVA, but to have full support for the informative hypothesis each contrast must be statistically significant.

V. *Linear contrast test* (*one-sided*)

For hypotheses imposing a simple order on a sequence of means, the linear contrast is a close approximation. The linear contrast weights for $k = 3$, 4, and 6 means (the values that we will use in the simulations) are:

$C_{\text{lin},3}$:	−1	0	+1			
$C_{\text{lin},4}$:	−3	−1	+1	+3		
$C_{\text{lin},6}$:	−5	−3	−1	+1	+3	+5

This provides, for example, $C_{\text{lin},4} = -3*M_1 - 1*M_2 + 1*M_3 + 3*M_4$. Since the contrast weights that are assigned to the sample means are increasing from negative to positive values, the value for $C_{\text{lin},k}$ will be positive

if the means are in the hypothesized order. Consider, for instance means $M_1 = 1$, $M_2 = 2$, $M_3 = 3$ and $M_4 = 4$ leading to $C_{\text{lin},4} = -3-2+3+12 = 10$, while $M_1 = 4$, $M_2 = 3$, $M_3 = 2$ and $M_4 = 1$ will provide $C_{\text{lin},4} = -12-3+2+3 = -10$. Therefore, $H_0 : C_{\text{lin},k} = 0$ is tested against $H_A : C_{\text{lin},k} > 0$ (*i.e.*, with one sided p-values). An advantage of this approach, compared to the previous, is that with one test the, for the hypothesis at hand, relevant p-value is obtained. A disadvantage is that the hypothesis that is tested (is the *linear* increase significantly different from zero?) is not equal to the originally stated hypotheses (are all means ordered from smallest to largest?).

VI. *Bayesian approach developed specifically for the evaluation of informative hypotheses*

Another method that will be evaluated is a Bayesian procedure specifically designed for the evaluation of informative hypotheses (see, for instance, [3] and [7]). With this model selection approach the support in the data for any hypothesis of interest is quantified with so-called Bayesian probabilities. Bayesian probabilities are numbers between zero and one reflecting the relative support for each hypothesis in a predefined set. In the simulation studies where the main interest is in a specific order constrained hypothesis, the set of models that will be compared consists of the null hypothesis $(H_0 : \mu_1 = \cdots = \mu_k)$ stating that all means are equal and the informative hypothesis imposing the ordering $(H_1 : \mu_1 < \cdots < \mu_k)$. To be able to compare the performance of the Bayesian model selection with the results based on *p*-values, in the simulation studies we will use the Bayesian probabilities to make dichotomous decisions, that is, either the informative hypothesis received the most support, or not. Note that making such dichotomous decisions does fit in the NHT framework (a result is statistically significant or not, usually judged with the 0.05 criterion) but not in the Bayesian framework, where it is up to the researcher to decide if he/she considers the resulting support for a certain hypothesis worthwhile ([3], page 51). A short summary of the Bayesian approach used in the simulations is provided in Appendix B. More extensive, non-technical introductions of Bayesian evaluation of informative hypotheses are provided in [8]-[10].

2.3. Simulation Studies

2.3.1. Defining the Populations

We investigated hypotheses expressing a simple ordering of $k = 3$, 4 and 6 means. The population parameter values were defined in agreement with the informative hypothesis and varied to obtain different effect sizes. In the context of an ANOVA the common effect size (ES) measure is Cohen's f, which is the ratio of the between groups standard deviation (σ_M) and the within groups (residual) standard deviation (σ_W). Cohen proposed to label $f = 0.1$ as a small effect, $f = 0.25$ as a medium effect and $f = 0.4$ as a large effect. In **Table 1**, the subpopulation means for each combination of k and ES are provided assuming residual variation $\sigma_W = 1$.

2.3.2. Power and Sample Sizes

From each population presented in **Table 1**, 10,000 data sets were sampled and subsequently analyzed with approaches I-V. The results for the Bayesian approach are based on 1000 data sets due to its intensive computation time. The sample sizes of the data sets are based on a power analysis using the following assumptions: 1) nowadays, it is more or less standard practice to start a research project with a power analysis to determine the number

Table 1. Population parameter values used for the simulation studies.

k	ES	Cohen's f	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6
3	Small	0.10	0	0.125	0.25			
	Medium	0.25	0	0.31	0.62			
	Large	0.40	0	0.49	0.98			
4	Small	0.10	0	0.09	0.18	0.27		
	Medium	0.25	0	0.225	0.45	0.675		
	Large	0.40	0	0.36	0.72	1.08		
6	Small	0.10	0	0.06	0.12	0.18	0.24	0.3
	Medium	0.25	0	0.145	0.29	0.435	0.58	0.725
	Large	0.40	0	0.235	0.47	0.705	0.94	1.175

of required participants to obtain power of 0.80; 2) we expect that most researchers perform their power analysis for the omnibus test (*i.e.*, for the one-way ANOVA) and that they do not take possible follow-up analyses and/or alpha corrections for multiple testing into account in the power analysis. Therefore, for each population, the required sample size to have 0.80 power for the omnibus ANOVA was determined and used in the simulations (numbers are reported in **Table 2**).

Additionally, for each of the six approaches, the approximate sample sizes required to obtain 0.80 *full support power*, as defined by the approach at hand and for the informative hypothesis of interest, are determined.

2.3.3. Results

In **Table 2**, the results for the six approaches are presented. The sample sizes used are provided in the first column and are based on a power analysis to obtain 0.80 power for the omnibus ANOVA. Note that all reported sample sizes are group sizes (N_j for $j = 1, \cdots, k$). The last six columns present the power to find full support for the research hypothesis with each of these approaches and using the sample sizes from the first column. So, for the 10,000 (1000 for approach VI) data sets that were sampled from prespecified populations with ordered means and effect sizes as specified, the resulting numbers in the table represent the proportions of these samples in which full support, as defined by each of the methods, for the hypothesis was found.

The results in column I show that, even if the only requirement is that the observed sample means should be in the hypothesized order, the power to find full support diminishes fast with an increasing number of means in the ordered hypothesis (approximately 0.70 for k = 3, 0.50 for k = 4, and 0.10 for k = 6). The power to find full support for the true order using the requirement that additional pairwise tests (one-sided or two-sided, and, with or without alpha corrections) should be significant reduces the power to zero in most cases (only for k = 3, full support power ranges between 0.02 - 0.15; see columns II-IV). Stated differently, with 10,000 replications from the same population, the true effect in the population was, with these methods, never fully confirmed.

The last two columns show that with the two confirmatory methods that do not rely on multiple tests, the power to find support for the ordered hypothesis is in almost all cases higher than the power of the omnibus test (ranging from 0.74 to 0.98). Further, it shows that for small effect sizes the linear contrast test has higher power than the Bayesian model selection approach, but that for medium and large effect sizes this is the other way around.

In **Table 3**, for the six approaches, the approximate required group sample sizes to obtain 0.80 full support power are provided. The numbers are obtained by running a sequence of simulations for each population (*i.e.*, combination of k and ES) with increasing sample sizes. We did not evaluate group sample sizes larger than 1000 because we believe they are not realistic in experimental research, so further precision seems unnecessary. The notation ">1000" in **Table 3** therefore means that with $N_j = 1000$ the full support power was still below 80%.

Table 2. Full support power for the six approaches for group sample sizes N_j that provide 0.80 power for the omnibus ANOVA (for several number of groups k and effect sizes ES).

k	ES	N_j^*	Approach					
			I	II	III	IV	V	VI
3	Small	310	0.72	0.05	0.02	0.14	0.87	0.79
	Medium	52	0.73	0.05	0.02	0.14	0.87	0.93
	Large	22	0.75	0.07	0.03	0.15	0.90	0.97
4	Small	271	0.50	0.00	0.00	0.00	0.91	0.74
	Medium	45	0.51	0.00	0.00	0.00	0.91	0.93
	Large	18	0.51	0.00	0.00	0.00	0.92	0.98
6	Small	205	0.10	0.00	0.00	0.00	0.95	0.76
	Medium	36	0.11	0.00	0.00	0.00	0.95	0.91
	Large	15	0.12	0.00	0.00	0.00	0.96	0.98

*Sample sizes N_j that provide 0.80 power for the ANOVA were determined using Gpower 3.1 [11].

Table 3. Approximate required group sample sizes N_j to obtain 0.80 full support power for each of the six approaches.

k	ES	Approach					
		I	II	III	IV	V	VI
3	Small	360	>1000	>1000	>1000	260	315
	Medium	60	220	260	180	45	26
	Large	25	90	110	75	18	6
4	Small	550	>1000	>1000	>1000	200	300
	Medium	90	470	600	390	32	23
	Large	35	190	240	150	12	4
6	Small	>1000	>1000	>1000	>1000	125	220
	Medium	280	>1000	>1000	>1000	22	21
	Large	110	500	670	415	9	4

The results show that huge samples are required to have reasonable full support power to detect a small ES with any of the approaches I-IV (ranging from 360 to >1000 per group). Approach V is, for small ES, most powerful (N_j range: 125 - 260) and outperforms the Bayesian approach (N_j range: 220 - 315). However, given that the smallest required N_j is still more than 100 respondents per subgroup, the results most of all show how difficult it is to find reliable and replicable support for specific expectations given that effect sizes are small.

The required sample sizes to find full support for medium effect sizes vary greatly between the approaches as well as between different numbers of subgroups. For $k = 3$, approaches I-IV require sample sizes ranging from $N_j = 60$ to 260, for $k = 4$ this increases to a range of $N_j = 90$ to 600 and for $k = 6$ only approach I gives a number below 1000 (*i.e.*, $N_j = 280$). Here, the gain of using approaches V or VI is clearly observed. Required N_j range from 21 to 45, where the Bayesian approach slightly outperforms the linear contrast approach. Similar patterns were observed for large effect sizes, where N_j ranged from 25 to 670 for approaches I-IV and from 4 to 18 for approaches V and VI.

Overall, the numbers in **Table 2** and **Table 3** clearly show that confirmatory methods that do not suffer from multiple testing issues (that is, approaches V and VI) are needed to have a good chance—with feasible sample sizes—to find full support for the true order of the means.

2.3.4. Additional Results for the Bayesian Approach

The Bayesian analysis for comparing H_0 with H_1 provides a so-called Bayes factor (BF): BF_{10} expresses how much more support the data provide for H_1 compared to H_0. Therefore, $BF_{10} > 1$ means more support for H_1, $BF_{10} = 1$ implies equal support for both hypotheses, and $BF_{10} < 1$ means more support for H_0.

In **Table 2**, the reported proportions (the "power" of the Bayesian approach) were based on counting how often, in 1000 replications, BF_{10} was bigger than 1. The interpretation of BFs is however not intended to be dichotomous ("hypothesis is supported or not"). To elaborate on the amount of support for the informative hypotheses that was found in the simulations, one could use the rules of thumb as presented by [12]. They propose that BF_{10} below 3 is still "not worth more than a bare mention", but that support can be claimed in the range 3 to 20 and that this support can be labeled as strong for $BF_{10} > 20$. For the simulations presented in **Table 2**, these elaborated results are provided in **Table 4** for the medium effect size.

From **Table 4**, we can see that in 7% to 9% of the samples the null hypothesis is favored over the ordered hypothesis, leading to a wrong conclusion. Note that this information was also presented in **Table 2**, where the "power of the Bayesian approach" was defined as finding $BF_{10} > 1$. On the second line of **Table 4**, we see that in about 10% of the samples the evidence is weakly in favor of the ordered hypothesis. In the remaining samples the support for the ordered hypothesis is substantial (in 22% $(k = 3)$ to 33% $(k = 6)$ of the samples) or even strong (49% for $k = 6$ to 62% for $k = 3$).

2.3.5. Non-Linear True Effects

A hypothesis stating a simple ordering of means is not equal to a hypothesis stating a non-zero linear effect. It is interesting to see if the power of approach V also holds when the population means are ordered from small to large, but not linearly, and how this power compares to approach VI that explicitly states the expected order.

A small simulation study was performed for k = 3, 4, 6, with $N_j = 25$, 50 for all cells and different non-linearly increasing population means. In **Table 5**, the investigated means are provided. The residual variance $\sigma_W = 1$ and this provides effect sizes f as reported in the second column of the table. The results are based on 10,000 samples for approach V and 1000 samples for approach VI and reported in the last two columns of **Table 5**.

The results show that the power of the Bayesian approach is higher for k = 3 and k = 4 and that the differences between approaches V and VI are largest for $N_j = 25$. For k = 6, approach V outperforms approach VI for $N_j = 50$. No clear pattern emerges for $N_j = 25$: in some cases the power of approach V is higher and in others it is the other way around.

3. An Illustration of a Two-Way Analysis of Variance

Often, ANOVA tests are done in the context of factorial designs, that is, with two or more factors and an interest in main and/or interaction effects. The example provided in the introduction will be used as an illustration. The researchers investigated stereotypes and gender differences in math performance in three subsequent studies [2]. In their first study, a group of highly selected respondents (see [2] for details) consisting of 28 men and 28 women, was randomized over easier and difficult math tasks. The goal of this study was to investigate if a specifically described expected interaction pattern was found. They formulated their expectation, for the studied population, as: "women underperform on difficult tests but perform just as well on easier test" ([2], page 9).

The hypothesized outcome, assuming general lower performance on the difficult test compared to the simple test, is represented in **Figure 1**. Formulated as an informative hypothesis, the expectation is:
$H_i : \mu_{w,d} < \mu_{m,d} < \{\mu_{w,s} = \mu_{m,s}\}$.

The tests executed and reported in [2], however, not directly address this expectation. They report F-tests for two main effects and an interaction effect (all $p < 0.05$), as well as posthoc pairwise comparisons of means. Therefore, multiple tests were required to come to the conclusion that, indeed, their expectation was supported.

A simulation study was designed to investigate the power to find full support for this specific expectation assuming different effect sizes and using several different approaches for the evaluation of the hypothesis. In Section 3.1 the approaches are presented and in Section 3.2 the design and results of the simulation study are provided.

Table 4. Proportions of different Bayes factors for k = 3, 4, 6 groups, medium effect size, and group sample sizes N_j providing power of 0.80 for the omnibus ANOVA.

BF_{10}	$k = 3/N_j = 52$	$k = 4/N_j = 45$	$k = 6/N_j = 36$
Less than 1	0.07	0.07	0.09
1 to 3	0.10	0.09	0.10
3 to 20	0.22	0.25	0.33
Greater than 20	0.62	0.60	0.49

Table 5. Comparison of the power of approaches V and VI when population means are increasing non-linearly.

	Cohen's f	μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	$N_j = 25$		$N_j = 50$	
								V	VI	V	VI
k = 3	0.25	0	0.2	0.6				0.54	0.80	0.85	0.91
	0.26	0	0.1	0.6				0.55	0.81	0.85	0.93
k = 4	0.25	0	0.1	0.5	0.6			0.69	0.82	0.94	0.94
	0.23	0	0.1	0.2	0.6			0.56	0.73	0.85	0.86
	0.22	0	0.25	0.35	0.6			0.55	0.71	0.85	0.87
k = 6	0.25	0	0.05	0.1	0.5	0.55	0.6	0.83	0.79	0.98	0.80
	0.20	0	0.05	0.1	0.15	0.20	0.6	0.54	0.59	0.83	0.74
	0.18	0	0.25	0.28	0.32	0.35	0.6	0.51	0.55	0.80	0.68

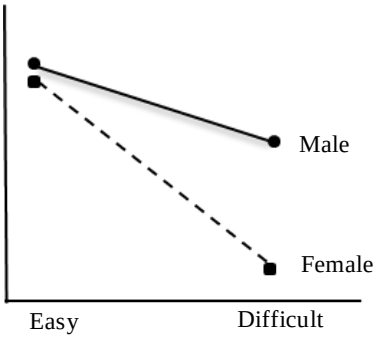

Figure 1. Hypothesized results for the study on stereotypes and gender differences in mathematics [2], with math performance on the y-axis.

3.1. Approaches

3.1.1. Factorial Approach

Most researchers would analyze these data with a two way ANOVA, testing for both main effects and the interaction effect. Different follow-up strategies could be considered, leading to three approaches described below. To limit the number of variations only results for two-sided tests and without alpha corrections are reported.

A. To conclude support for the theory, both main effects as well as the interaction effect should be statistically significant and the sample means (M_j) should be in a specific order, that is: $(M_{w,d} < M_{m,d})$, $(M_{m,d} < M_{m,s})$, $(M_{m,d} < M_{w,s})$.

B. To conclude support for the theory, in approach B the three omnibus tests should be significant and the sample means in the right order (as in A) but also the simple main effects should support the theory. This implies finding a significant result for the test $H_0 : \mu_{w,d} = \mu_{m,d}$ and a non-significant result for $H_0 : \mu_{w,s} = \mu_{m,s}$.

C. Following [2], as a follow-up to the requirements of approach A, we tested all pairwise comparisons of means. The results should be non-significant for $H_0 : \mu_{w,s} = \mu_{m,s}$, while the other 5 pairwise comparisons must be statistically significant.

3.1.2. One Way Approach

Since the factorial approach is rather exploratory (testing for *any* main effect and *any* interaction and not for the specific, expected patterns), the omnibus tests could be skipped and instead planned comparisons on the four subgroup means could be executed and interpreted. Note that this implies ignoring the factorial structure in the design. Two approaches are included in the simulations:

D. The first approach is based on planned comparisons on specific contrasts. The ordering $\mu_{w,d} < \mu_{m,d} < \{\mu_{w,s} = \mu_{m,s}\}$ is captured by: $C_1 = -3 * M_{w,d} - 1 * M_{m,d} + 2 * M_{w,s} + 2 * M_{m,s}$. Support for the expectation can be concluded if the test $H_0 : C_1 = 0$ against $H_1 : C_1 > 0$ (*i.e.*, with a one-sided *p*-value) is statistically significant. However, C_1 does not include the expectation that the last two means are not different. So, in addition we formulate the contrast: $C_2 = 0 * M_{w,d} + 0 * M_{m,d} + 1 * M_{w,s} - 1 * M_{m,s}$, and to conclude support for the theory this contrast test should *not* be significant (two-sided *p*-value).

E. The Bayesian approach for informative hypotheses can evaluate the expected pattern directly. In the simulation we will evaluate how often the informative hypothesis $H_{\text{Spencer}} : \mu_{w,d} < \mu_{m,d} < \{\mu_{w,s} = \mu_{m,s}\}$ receives more support than the null hypothesis $H_0 : \mu_{w,d} = \mu_{m,d} = \mu_{w,s} = \mu_{m,s}$. Note again that, in a Bayesian framework, one would not make dichotomous decisions but evaluate the amount of support for each hypothesis. Furthermore, H_i could be compared to several other alternative hypotheses than H_0, e.g., a hypothesis imposing no constraints $(H_A : \mu_{w,d}, \mu_{m,d}, \mu_{w,s}, \mu_{m,s})$ or a hypothesis expressing an alternative competing informative hypothesis. However, the current choices (evaluating against H_0 and drawing dichotomous conclusions) are made to be able to meaningfully compare the results with the NHT results in approaches A-D.

3.2. Simulation Study and Results

For the simulation study, several populations were specified with means in agreement with the informative hypothesis of [2]. The residual variance was always one, and the differences between the means were increasing form relatively small to larger differences. The population means in five simulations are presented in **Table 6**. Results are also found in the table and are based on drawing 10,000 (1000 for approach E) samples with a sample size of 50 per group.

The results show, once again, that the power to find full support for a specific expectation dramatically decreases when several multiple tests are involved, as in approach A, B and—most of all—C. The power of approach D is already much higher. This can be explained by the fact that, here, only two tests were involved of which the first contrast specifically represents the order of interest and was evaluated with a one-sided p-value, that is, it was evaluated in a relatively powerful, confirmatory way. Finally, the Bayesian approach (E) slightly outperforms approach D. The advantage of specifying precisely what one wants to know and evaluating this with a direct approach results in the highest power levels.

4. Discussion

Attention for limitations of null hypothesis testing in general, e.g., [13]-[15], and problems with power, lack of replication, and multiple testing specifically, e.g., [4] [16], is widespread in both statistical and applied research literature. In the past two decades, a Bayesian approach for the evaluation of informative hypothesis was presented as an alternative, confirmatory approach, e.g., [7] [17] [18]. In these papers it is often claimed that with the formulation and evaluation of informative hypotheses more powerful methods are obtained. In a few papers, some examples are provided to support this claim with numbers, e.g., [19] [20]. However, so far, no systematic study of the power of different—exploratory and confirmatory—approaches was reported and this is, therefore, the main contribution of this paper.

We presented several simulations in the context of evaluating a simple ordering of k means in a one-way design. The results, however, present a more general message and are similar when the k means come from a factorial design and irrespective of which expected pattern of means is evaluated. To illustrate this, one example of a two-way analysis of variance and an informative hypothesis that did impose a different set of constraints on the means was also provided.

Results in this paper show that the approaches that are mostly found in the research literature, that is, analysis of variance omnibus tests with multiple follow-up comparisons of means, have very limited power to detect the true pattern of means. Approaches that are specifically designed for the evaluation of prespecified expectations like planned contrast testing or the Bayesian approach for informative hypotheses do much better. Typical differences observed in the simulations for the one-way design were power levels between 0% - 15% for approaches based on multiple testing, whereas the power of the confirmatory approaches reached power levels between 80% - 100%.

Additional simulations were done to investigate what sample sizes would be needed to have reasonable power with the commonly used approaches. The main conclusion from these simulations is that it is practically unfeasible to detect the true pattern of means with such approaches if the effect size is small, and that it still requires huge group sample sizes to detect medium effects (N_j between 180 and more than 1000) or large effects (N_j between 75 and 670). Again, the two confirmatory approaches fare much better although, also here, the sample sizes to detect small effect sizes are relatively large (between 125 - 315 per group).

Table 6. Comparison of the power of approaches A-E when population means are in agreement with the hypothesis of interest with increasing differences between the means (N_j = 50).

$\mu_{w,d}$	$\mu_{m,d}$	$\mu_{w,s}$	$\mu_{m,s}$	Approach				
				A	B	C	D	E
0	0.05	0.1	0.1	0.00	0.00	0.00	0.08	0.09
0	0.1	0.2	0.2	0.00	0.00	0.00	0.21	0.24
0	0.2	0.4	0.4	0.00	0.00	0.00	0.62	0.67
0	0.3	0.6	0.6	0.02	0.02	0.00	0.91	0.96
0	0.5	1.0	1.0	0.17	0.17	0.06	0.97	1.0

The results lead to a couple of recommendations. First of all, if specific expectations are formulated beforehand, as is, for instance, often the case in psychological experiments like those described in the paper, we strongly recommend considering approaches that use as few multiple tests as possible. Planned contrasts have much more power to detect the true patterns than omnibus ANOVA's with several follow-up tests, and so does the Bayesian approach. Whenever the expectation can be formulated in one contrast (e.g., the linear contrast we used in the one-way design to reflect the simple ordering of means), the differences in power between the contrast testing approach and the Bayesian approach are negligible for the effect sizes and numbers of groups investigated in this paper. A potential advantage of the Bayesian approach is flexibility in terms of the types of hypotheses that can be formulated and evaluated. Any expected pattern that can be expressed using a combination of smaller than (<), bigger than (>), equal to (=), and/or, no mutual constraint (,) can be evaluated using the Bayesian approach. An example where planned contrast testing required two tests and therefore resulted in less power to detect the true pattern than the Bayesian method was provided in the context of a factorial design and specific expectations about the interaction effect of the two factors.

The results of this paper also show how variable different samples from the same population can be. Although sampling variability is a concept known to all researchers that are familiar with data analysis and NHT, published literature shows that many researchers do often underestimate the size and consequences of sampling variability in their own study. Not finding a specific expected difference between two means, for instance, is often explained by substantive arguments. The fact that this could very likely be a type 2 error, due to limited power, is hardly ever mentioned, especially when some other interesting comparisons did reach statistical significance. Likewise, the finding of a significant difference between certain means that was not a priori expected often receives considerable attention. However, in the context of multiple testing the probability of finding at least one significant difference is large and therefore it might just as well be a chance finding (inflated type 1 error due to multiple testing). It seems that, significant and non-significant results are too often interpreted as rather certain indicators of the true effects. With this paper, we hope to contribute to the awareness that results from a single study can only provide conclusions with very limited certainty and that replication studies are crucial.

Another recommendation relates to the publication process, where a clearer distinction could be made between confirmatory and exploratory analyses. When specific theories or expectations are specified a priori and confirmatory methods are used to evaluate the expectations, conclusions can be relatively strong, although the need for replication studies will always remain. Other findings, or when no specific hypotheses were formulated, should be reported acknowledging the exploratory nature of the results. One can conclude that interesting findings were seen in this particular data set but replications with new data, and preferably confirmatory methods, are required before it can be concluded if they reflect real effects or chance findings.

5. Conclusion

The need for replication leads to a final recommendation, which is not at all original, but crucial to the accumulation of scientific knowledge. All results of all tests within a study should be properly reported, irrespective of their statistical significance, including the appropriate sample statistics (e.g. means, standard deviations, group sample sizes). This holds not only for non-significant findings within a study but also for studies where no significant results were found at all and that are currently often hard to get published. Only when these types of publication bias are avoided, replication studies can be properly synthesized and results judged for what they are worth. If such a publication culture could be established we can work towards accumulation of knowledge in a truly scientific way.

References

[1] Van den Hout, M.A., Rijkeboer, M.M., Engelhard, I.M., Klugkist, I., Hornsveld, H., Toffolo, M. and Cath, D.C. (2012) Tones Inferior to Eye Movements in the EMDR Treatment of PTSD. *Behaviour Research and Therapy*, **50**, 275-279. http://dx.doi.org/10.1016/j.brat.2012.02.001

[2] Spencer, S.J., Steele, C.M. and Quinn, D.M. (1999) Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, **35**, 4-28. http://dx.doi.org/10.1006/jesp.1998.1373

[3] Hoijtink, H. (2012) Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists. Chapman and Hall/CRC, London.

[4] Maxwell, S.E. (2004) The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences,

and Remedies. *Psychological Methods*, **9**, 147-163. http://dx.doi.org/10.1037/1082-989X.9.2.147

[5] Williams, L.J. and Abdi, H. (2010) Fisher's Least Significant Difference (LSD) Test. In: Salkind, N., Ed., *Encyclopedia of Research Design*, Sage, Thousand Oaks. http://dx.doi.org/10.4135/9781412961288.n154

[6] Rosenthal, R., Rosnow, R.L. and Rubin, D.B. (2000) Contrasts and Effect Sizes in Behavorial Research. A Correlation Approach. Cambridge University Press, Cambridge.

[7] Klugkist, I., Laudy, O. and Hoijtink, H. (2005) Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, **10**, 477-493. http://dx.doi.org/10.1037/1082-989X.10.4.477

[8] Béland, S., Klugkist, I., Raîche, G. and Magis, D. (2012) A Short Introduction into Bayesian Evaluation of Informative Hypotheses as an Alternative to Exploratory Comparisons of Multiple Group Means. *Tutorials in Quantitative Methods for Psychology*, **8**, 122-126.

[9] Klugkist, I., Van Wesel, F. and Bullens, J. (2011) Do We Know What We Test and Do We Test What We Want to Know? *International Journal of Behavioral Development*, **35**, 550-560. http://dx.doi.org/10.1177/0165025411425873

[10] Van de Schoot, R., Mulder, J., Hoijtink, H., van Aken, M.A.G., Dubas, J.S., de Castro, B.O., Meeus, W. and Romeijn, J.-W. (2011) An Introduction to Bayesian Model Selection for Evaluating Informative Hypotheses. *European Journal of Developmental Psychology*, **8**, 713-729. http://dx.doi.org/10.1080/17405629.2011.621799

[11] Faul, F., Erdfelder, E., Lang, A.G. and Buchner, A. (2007) G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods*, **39**, 175-191. http://dx.doi.org/10.3758/BF03193146

[12] Kass, R.E. and Raftery, A. (1995) Bayes Factors. *Journal of the American Statistical Association*, **90**, 773-795. http://dx.doi.org/10.1080/01621459.1995.10476572

[13] Nickerson, R.S. (2000) Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, **5**, 241-301. http://dx.doi.org/10.1037/1082-989X.5.2.241

[14] Cohen, J. (1994) The Earth Is Round ($p < .05$). *American Psychologist*, **49**, 997-1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

[15] Royall, R.M. (1997) Statistical Evidence. A Likelihood Paradigm. Chapman & Hall, New York.

[16] Cumming, G. (2008) Replication and *p* Intervals: *p* Values Predict the Future Only Vaguely, but Confidence Intervals Do Much Better. *Perspectives on Psychological Science*, **3**, 286-300. http://dx.doi.org/10.1111/j.1745-6924.2008.00079.x

[17] Mulder, J., Hoijtink, H. and 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. http://dx.doi.org/10.1016/j.jspi.2009.09.022

[18] Van Wesel, F., Hoijtink, H. and Klugkist, I. (2010) Choosing Priors for Inequality Constrained Normal Linear Models: Methods Based on Training Samples. *Scandinavian Journal of Statistics*, **38**, 666-690. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9469.2010.00719.x/full http://dx.doi.org/10.1111/j.1467-9469.2010.00719.x

[19] Kuiper, R.M. and Hoijtink, H. (2010) Comparisons of Means Using Exploratory and Confirmatory Approaches. *Psychological Methods*, **15**, 69-86. http://dx.doi.org/10.1037/a0018720

[20] Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M.A.G., Orobio de Castro, B., Meeus, W. and Romeijn, J.W. (2011) Evaluating Expectations about Negative Emotional States of Aggressive Boys Using Bayesian Model Selection. *Developmental Psychology*, **47**, 203-212. http://dx.doi.org/10.1037/a0020957

Appendix A: Tests Used in Approaches I-V (Section 2) and A-D (Section 3)

Notations used for one-way design:

$N =$ total sample size; $M =$ overall sample mean;
$k =$ number of groups $(j = 1, \cdots, k)$; $N_j =$ group sample size $(i = 1, \cdots, N_j)$;
$M_j =$ group sample mean; $S_j^2 =$ group sample variance.

Approach I

The ANOVA test result is determined using: $F = \dfrac{MS_{\text{between}}}{MS_{\text{within}}}$,

with:

$$MS_{\text{between}} = \frac{\sum_{j=1}^{k} N_j \left(M_j - M\right)^2}{df_{\text{between}}},$$

$$MS_{\text{within}} = \frac{\sum_{j=1}^{k} \left(N_j - 1\right) S_j^2}{\mathrm{d}f_{\text{within}}},$$

$$df_{\text{between}} = k - 1,$$

$$df_{\text{within}} = N - k,$$

and evaluated using the $F\left(df_{\text{between}}, df_{\text{within}}\right)$ distribution.

Approaches II and III

The pairwise t-test is based on: $T = \dfrac{\left|M_j - M_{j-1}\right|}{\sqrt{MS_{\text{within}}\left(\dfrac{1}{N_j} + \dfrac{1}{N_{j-1}}\right)}}$,

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution. For approach II the p-value is equal to the two-sided tail probability. For approach III, the p-value is that probability multiplied by $(k-1)$.

Approach IV

Each planned contrast t-test is based on: $T = \dfrac{M_j - M_{j-1}}{\sqrt{MS_{\text{within}}\left(\dfrac{1}{N_j} + \dfrac{1}{N_{j-1}}\right)}}$,

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution and the one-sided tail probability (taking the hypothesized order into account).

Approach V

Denoting the linear contrast weight for M_j with λ_j the test is based on:

$$T = \frac{C_{\text{lin},j}}{\sqrt{MS_{\text{within}} * \sum_j \dfrac{\lambda_j^2}{N_j}}},$$

with:

$C_{\text{lin},3} = -1 * M_1 + 0 * M_2 + 1 * M_3$,

$C_{\text{lin},4} = -3 * M_1 - 1 * M_2 + 1 * M_3 + 3 * M_4$,

$C_{\text{lin},6} = -5 * M_1 - 3 * M_2 - 1 * M_3 + 1 * M_4 + 3 * M_5 + 5 * M_6$,

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution and the one-sided tail probability (taking the hypothesized order into account).

Notations used for two-way design:

$N =$ total sample size; $M =$ overall sample mean;

$k =$ number of levels first factor $(j = 1, \cdots, k)$; $h =$ number of levels second factor $(g = 1, \cdots, h)$;
$N_{jg} =$ cell sample size $(i = 1, \cdots, N_{jg})$; $N_{j.}$ and $N_{.g} =$ marginal sample sizes;
$M_{jg} =$ cell sample mean; $M_{.g}$ and $M_{j.} =$ marginal sample means;
$S^2_{jg} =$ cell sample variance.

Approach A

The ANOVA test results are determined using: $F = \dfrac{MS_{\text{effect}}}{MS_{\text{within}}}$,

with:

$$MS_{\text{effect1}} = \frac{\sum_{j=1}^{k} N_{j.}\left(M_{j.} - M\right)^2}{df_{\text{effect1}}}, \quad df_{\text{effect1}} = k - 1, \text{ for main effect 1,}$$

$$MS_{\text{effect2}} = \frac{\sum_{g=1}^{h} N_{.g}\left(M_{.g} - M\right)^2}{df_{\text{effect2}}}, \quad df_{\text{effect2}} = h - 1, \text{ for main effect 2,}$$

$$MS_{\text{effect3}} = \frac{\sum_{j=1}^{k}\sum_{g=1}^{h} N_{jg}\left(M_{jg} - M_{j.} - M_{.g} + M\right)^2}{df_{\text{effect3}}}, \quad df_{\text{effect3}} = (k-1)*(h-1), \text{ for the interaction effect,}$$

$$MS_{\text{within}} = \frac{\sum_{j=1}^{k}\sum_{g=1}^{h}\left(N_{jg} - 1\right)S^2_{jg}}{df_{\text{within}}}, \quad df_{\text{within}} = N - k * h,$$

and evaluated using the $F\left(df_{\text{effect}}, df_{\text{within}}\right)$ distributions.

Approach B

Two *t*-tests are used for the simple main effects:

$$T = \frac{\left|M_{wd} - M_{md}\right|}{\sqrt{MS_{\text{within}}\left(\frac{1}{N_{wd}} + \frac{1}{N_{md}}\right)}} \quad \text{and} \quad T = \frac{\left|M_{ws} - M_{ms}\right|}{\sqrt{MS_{\text{within}}\left(\frac{1}{N_{ws}} + \frac{1}{N_{ms}}\right)}}$$

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution. Note that the first test is evaluated with the one-sided tail probability, whereas the second is not.

Approach C

All pairwise t-tests comparing group jg with $j'g'$ are based on:

$$T = \frac{\left|M_{jg} - M_{j'g'}\right|}{\sqrt{MS_{\text{within}}\left(\frac{1}{N_{jg}} + \frac{1}{N_{j'g'}}\right)}}$$

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution.

Approach D

Denoting the contrast weights for M_{jg} with λ_{jg} the tests used are:

$$T = \frac{C}{\sqrt{MS_{\text{within}} * \sum_{j}\sum_{g}\frac{\lambda^2_{jg}}{N_{jg}}}}$$

and evaluated using the $T\left(df_{\text{within}}\right)$ distribution. Note that the first test with $C_1 = -3 * M_{w,d} - 1 * M_{m,d} + 2 * M_{w,s} + 2 * M_{m,s}$ is evaluated with the one-sided tail probability, whereas the second test with $C_2 = 0 * M_{w,d} + 0 * M_{m,d} + 1 * M_{w,s} - 1 * M_{m,s}$ is not.

Appendix B: A Short Summary of the Bayesian Method in Approach VI (Section 2) and E (Section 3)

Note that this appendix was previously published as an appendix to the paper by Klugkist, Van Wesel, Bullens (2011).

Here, we will shortly outline the Bayesian approach applied in this paper for the analysis of variance (ANOVA) model:

$$y_i = \sum_{j=1}^{J} \mu_j d_{ji} + \varepsilon_i$$

where y_i is the outcome of person i $(i = 1, \cdots, n)$, d_{ji} denotes group membership for $j = 1, \cdots, J$ groups (where $d_{ji} = 1$ if person is a member of group j, and zero otherwise), and μ_j is the mean of group j. The residuals ε_i are assumed to be independent and normally distributed with mean zero and variance σ^2.

The ANOVA model has the following likelihood:

$$f\left(y \middle| D, \mu, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_i - \left[\sum_{j=1}^{J} \mu_j d_{ji} \right] \right)^2 \right\}$$

where $y = \{y_1, \cdots, y_n\}$, $D = \{d_1, \cdots, d_J\}$, $d_j = \{d_{j1}, \cdots, d_{jn}\}$, and $\mu = \{\mu_1, \cdots, \mu_J\}$.

Prior based on training data

The method and software used for the Bayesian analysis is based on Van Wesel, Hoijtink and Klugkist, (2011). In this paper, a thorough investigation of different priors that can be used for the analysis of informative hypotheses (H_i) in the context of ANOVA is presented. The method is based on the use of an encompassing prior, that is, a (low informative) prior is specified for the unconstrained hypothesis H_A and the prior distributions for the constrained hypotheses can be derived by truncation of the prior parameter space, using:

$$g\left(\mu, \sigma^2 \middle| H_i\right) = \frac{g\left(\mu, \sigma^2 \middle| H_i\right) I_{H_i}}{\int g\left(\mu, \sigma^2 \middle| H_i\right) I_{H_i} \mathrm{d}\mu \mathrm{d}\sigma^2}$$

where I_{H_i} is an indicator function with value one if the means are in agreement with H_i, and zero otherwise.

The specification of the unconstrained prior $g\left(\mu, \sigma^2 \middle| H_i\right)$ is based on training data (Berger and Pericchi, 2004, 1996; Perez and Berger, 2002). A training sample is a small part of the data that can be used to update the reference prior for the ANOVA model, $1/\sigma^2$ (Bernardo, 1979), such that the resulting posterior is proper but also low informative and objective (*i.e.*, no subjective information is used). In the approaches described in the references above, multiple training samples are used and the results are combined in different ways. Van Wesel *et al.* (2011) proposed a prior that is based on the same principles but tailored for constrained hypotheses and less computer intensive (*i.e.*, faster). This prior is called the average constrained posterior prior (ACPP). For a detailed explanation and elaborate motivation for this prior we refer to the original paper.

The general form of the ACPP is:

$$g^{\mathrm{ACPP}}\left(\mu, \sigma^2 \middle| H_A\right) = N\left(\mu \middle| \hat{\mu}, \hat{\Sigma}\right) \times \mathrm{Inv} - \chi^2\left(\sigma^2 \middle| \hat{\nu}, \hat{\kappa}^2\right)$$

where $N(\cdot)$ denotes the multivariate normal distribution with a mean parameter and covariance matrix, and $\mathrm{Inv} - \chi^2(\cdot)$ denotes the scaled inverse chi-square distribution with the degrees of freedom and a scale parameter.

Posterior

The posterior distribution based on the ACPP is:

$$h^{\mathrm{ACPP}}\left(\mu, \sigma^2 \middle| y, D, H_A\right) \propto f\left(y \middle| D, \mu, \sigma^2\right) \times N\left(\mu \middle| \hat{\mu}, \hat{\Sigma}\right) \times \mathrm{Inv} - \chi^2\left(\sigma^2 \middle| \hat{\nu}, \hat{\kappa}^2\right)$$

Bayes factors

The Bayes factor comparing two hypotheses is the ratio of two marginal likelihoods. A marginal likelihood, for instance $m\left(y \middle| H_A\right)$, is the density of the data averaged over the prior distribution of H_A. Chib (1995) noted that for the estimation of the marginal likelihood it can be useful to use the expression (imputing our choice of prior and subsequent posterior):

$$m\left(y\middle|H_A\right)=\frac{f\left(y\middle|D,\mu,\sigma^2\right)g^{\text{ACPP}}\left(\mu,\sigma^2\middle|H_A\right)}{h^{\text{ACPP}}\left(\mu,\sigma^2\middle|y,D,H_A\right)}.$$

Subsequently, Klugkist and Hoijtink (2007) derived that in the context of encompassing priors (*i.e.*, the constrained model is nested in the unconstrained), the Bayes factor comparing an informative hypothesis H_i with the unconstrained H_A $\left(\text{BF}_{i,A}\right)$ reduces to the ratio of two proportions: the proportion of the unconstrained posterior distribution in agreement with the constraints of H_i, and the proportion of the unconstrained prior distribution in agreement with the constraints of H_i. These proportions are estimated using (MCMC) sampling methods.

Note that in this approach each H_i is evaluated against H_A but that mutual comparison of, for instance H_{i1} with H_{i2} is also possible, using:

$$\text{BF}_{i1,i2}=\frac{\text{BF}_{i1,A}}{\text{BF}_{i2,A}}.$$

Posterior model probabilities

Using a uniform prior on the model space, the posterior model probabilities for t $\left(t=1,\cdots,T\right)$ hypotheses are computed using:

$$\text{PMP}\left(H_t\right)=\frac{\text{BF}_{t,A}}{\sum_{t=1}^{T}\text{BF}_{t,A}}.$$