Fayette Klaassen

The latest update on Bayesian informative hypothesis testing

The latest update on Bayesian informative hypothesis testing

De nieuwste update over Bayesiaanse informatieve hypothese toetsing
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 10 januari 2020 des middags om 2.30 uur

door

Fayette Klaassen

geboren op 13 mei 1992
te Amsterdam

Promotoren:

Prof.dr. H.J.A. Hoijtink

Prof.dr. I.G. Klugkist

The studies in this thesis were funded by the Netherlands Organization for Scientific Research (project 406-12-001).

Beoordelingscommissie:

Dr.ir. J. Mulder

Prof.dr. A. Postma

Prof.dr. A.G.J. van de Schoot

Prof.dr. J.K. Vermunt

Prof.dr. E.M. Wagenmakers

The latest update on Bayesian informative hypothesis testing.
Proefschrift Universiteit Utrecht, Utrecht.
Met lit. opg. - Met samenvatting in het Nederlands.

Cover design: Fayette Klaassen
Cover DTP: Mirjam van Laar
Print: Ridderprint | www.ridderprint.nl

Contents

Chapter 1

Introduction

"Dad is a man, because he can whistle and only men can whistle."

My parents kept a notebook where they recorded memorable phrases of my and my sisters when we were growing up. I was six years old when I claimed the above, believing whistling abilities determined gender. This inference was informed by the observations I had made throughout my life. The only people that I had ever heard whistle were my father and the 'whistling neighbor' who we heard all summer long from his backyard. My mother never whistled, nor had either of my sisters ever demonstrated any whistline abilities. My experience told me that only men could whistle and that this was a unique differentiating feature between men and women. Investigative as I was, this theory was tested by trying to mimic the general facial movements I had observed. I put my lips together, sucked in my cheeks and blew air out. Without success, I spattered and exhaled air. My conclusion: those who *can* whistle are men, those who cannot are women.

1.1 The research cycle

Throughout our lives we formulate theories, collect evidence and update our knowledge continuously. The theory that only men could whistle resulted in testable expectations, experimenting and data collection, evidence and finally a conclusion. The data confirmed my theory, but I did not think of competing theories or whether observing two successes and four failures was sufficient to confirm the theory. In fact, this particular theory only required one observation to discard my gender-defining-whistle theory: My kindergarten teacher, a woman, whistled. With this new knowledge I realized there was something else that distinguished whistlers from non-whistlers. Was it the ability to get the right shape of lips and tonge? Other new questions developed. What was the amount force required to make a difference between air and sound? How do you change notes so you can whistle a melody? Even though my initial theory was rejected, over the years my knowledge, theories and questions on whistling-abilities keep increasing.

Six-year old me passed through various stages of the research cycle. Some twenty years later, the research of four years presented in this dissertation touches upon the research cycle

once more. At age six, I unknowingly formulated hypotheses based on expectations from a personal theory. I was satisfied with six observations that all confirmed this particular theory. At age six, I unknowingly drew conclusions about whether a phenomenon (whistling-ability) was present for all men based on the observation of a few cases. At age six, I unknowingly performed Bayesian updating and continuously learned about, developed and rejected many more theories. Similar to the whistling example, I believed women could bake pancakes and braid hair while men could not. I updated my theories when I ate pancakes baked by my dad and when no apple tree would grow in my stomach after swallowing the seeds. I did learn how to whistle and though I know men *can* braid hair but doubt the quality. Mostly, I learned to be critical of my assumptions and ask questions. Some twenty years later I investigated the value and importance of these unknowingly taken research steps. This dissertation contains my latest updates about Bayesian informative hypothesis testing.

1.2 Bayesian informative hypothesis evaluation

The title of this dissertation, *The latest update on Bayesian informative hypothesis testing*, refers to its two main topics: *informative hypothesis evaluation* and *Bayesian updating*. This section provides a brief introduction to these topics and how they are connected for a general framework. Each chapter contains the necessary definitions on these concps In order to not needlessly repeat definitions and concepts, the more elaborate definitions are available in the relevant chapters.

1.2.1 Bayesian statistics

Bayesian statistics differs from classical or frequentist statistics in the part of an analysis that is considered random. Data are analyzed through a statistical model, that is, a collection of parameters that explain the observed data. In frequentist statistics the probability of data given a statistical model is determined, e.g. the probability of observing a woman who whistles, given that only men can whistle. In contrast, Bayesian statistics determines the probability of a statistical model given the data, e.g. the probability that only men can whistle, given that you observe a woman who whistles. Bayes' theorem shows how knowledge about this probability of interest is updated:

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(B)} \qquad \text{(Bayes' theorem)} \qquad (1.1)$$

The theorem describes how the probability of A conditional on B, ($P(A \mid B)$can be obtained as a function of $P(A)$, $P(B \mid A)$ and $P(B)$. In other words, it describes how we can obtain the conditional probability of A given B, for example a hypothesis A given a dataset B, using the probabilities of the data B, hypothesis A and the probability of the data given the hypothesis $P(B \mid A)$.

1.2.2 Informative hypothesis testing

Hypotheses are the link between the observed data and the deveped theories. Data are used to draw conclusions about theories by statistically testing hypotheses. In the classical null hypothesis significance testing (NHST) a null and an alternative hypothesis H_0 and H_a are considered, that is, the expectation of *no* effect (H_0) or *some* effect H_a. Many arguments have been made against the use of this procedure of hypothesis testing, since both H_0 and H_a do not appear to reflect a researchers expectations (e.g., Cohen, 1994; Klugkist, Wesel, & Bullens, 2011; van de Schoot, Hoijtink, & Romeijn, 2011; Wagenmakers, 2007). Rarely is the exact absence of an effect expected to be true in the population or is the interest of a researcher to confirm the presence of *any* effect. Specific expectations are often evaluated by means of a posteriori analyses, rather than directly. Constrained hypothesis testing has been introduced as a way to more directly evaluate the expectations of researchers. Specifically, Klugkist, Laudy, & Hoijtink (2005), Hoijtink, Klugkist, & Boelen (2008), Mulder et al. (2009), Hoijtink (2012) developed straighforward methods to evaluate these constrained hypotheses by means of Bayesian hypothesis tests.

1.2.3 Bayesian hypothesis testing

Bayesian hypothesis testing allows for the specification of any interesting model and the comparison of multiple theories. Both the use of Bayesian hypothesis testing with a Bayes factor and the use of informative hypotheses have gained in popularity over the past years (Mulder & Wagenmakers, 2016). Easy to use software with options for Bayesian analyses or the inclusion of constrained hypotheses is more accessible (e.g. JASP Team, 2018; or R packages such as Morey & Rouder, 2018; Gu, Hoijtink, Mulder, & Lissa, 2019; Merkle & Rosseel, 2018). Consequently, the share of research in the social and behavioral sciences that adopts Bayesian informative hypothesis testing is slowly increasing. Equation 1.1 tells us how we can obtain the probability of a single hypothesis. Knowing the probability of one hypothesis is not particularly useful if the goal is to make a comparison between multiple hypotheses. By taking a ratio of Equation 1.1 for two different hypotheses we get:

$$\begin{aligned} \frac{P(A_1|B)}{P(A_2|B)} &= \frac{P(A_1)}{P(A_2)} \times \frac{P(B|A_1)}{P(B|A_2)} \\ \text{posterior odds} &= \text{prior odds} \times \text{Bayes factor} \end{aligned} \tag{1.2}$$

Equation 1.2 shows how posterior odds, the ratio of two posterior probabilities, are obtained by updating the prior odds, a ratio of two prior probabilities, with the Bayes factor, the ratio of two marginal distributions (Kass & Raftery, 1995). This Bayes factor is the rate with which the prior odds are updated, also referred to as the *evidence* (Morey, Romeijn, & Rouder, 2016).

1.2.4 Updating

Both the formulation of hypotheses and the computation of evidence for these hypotheses can continuously be done. A single step of data collection and analysis gives a temporary answer to the question at hand. The updating of prior to posterior knowledge in Equation

1.1 and 1.2 can be repeated as more data becomes available. New theories can be developed, old hypotheses can become outdated. Equation 1.1 shows that knowledge is never final and with each new piece of information, we learn more about a puzzle with an unknown number of pieces. All we can do is look at the latest update, and see where we are right now.

1.3 The latest update

With the increased use of Bayesian informative hypothesis testing, practical, philosophical and methodological questions arise. One of the arguments in favor of a Bayes factor as opposed to a p-value is the lack of an arbitrary cut-off for making decisions (Wagenmakers, 2007). The Bayes factor has a meaningful interpretation, namely the rate with which knowledge is updated. However, practice shows that guidelines and corresponding cut-offs are developed (e.g. Kass & Raftery, 1995; Wagenmakers, Wetzels, Borsboom, & Maas, 2011). This results in questions: what conclusions can you draw with a Bayes factor and what does a particular level of evidence mean in practical terms? With the possibility of evaluating any hypothesis one can think of, comes the problem of choosing the relevant hypotheses. Which comparisons are interesting and important to make and how can you make a valuable comparison? The importance of prior distribution and prior probability specification in Bayesian hypothesis testing needs to be discussed and comes with great responsibility. Every decision in a Bayesian hypothesis test, much like in a frequentist hypothesis test, needs to be justified and with little established practice, many ideas exist about the 'right' approach. This dissertation addresses a few of these questions.

One step in the research cycle is to collect data for hypothesis testing. The amount of data required to answer a research question depends on the value of making wrong conclusions. The link between sample size, power and error probabilities is well-researched in the NHST framework. In Bayesian statistics research this relationship is less discussed and the value of power and unconditional error probabilities are debated. Chapter 2 presents four sample size determination methods for informative hypothesis testing by means of Bayes factors. The value of power and (un)conditional error probabilities and their link with sample size for Bayesian hypothesis tests are discussed.

Another step in the research cycle is to translate the results from a statistical analysis into a conclusion. The analysis should match the research question to provide a sensible conclusion. Many hypothesis tests concern the presence and direction of *population* effects. However, in practice the conclusions from these hypothesis tests often are at the *individual* level. For example, after analyzing the effectiveness of a medication in the population, it is prescribed to individuals. The average effect does not imply the medicine works for all individuals. In many situations the main interest is in the individual effects rather than population effects. Chapters 3 and 4 describe how Bayesian hypothesis testing can be used to synthesize the results from multiple individual analyses. Bayesian statistics can be used to continuously add data and sequentially update knowledge about population effects. This process is called *updating*. Alternatively, data from multiple individuals can be analyzed separately and combined to learn about how the homogeneity (similarity) of individual effects. Chapter 3 presents the methodology and Chapter 4 is a hands-on description for how to execute such an analysis. For Chapter 2 an R package has been developed, and

for Chapter 3 an R Shiny application has been developed. Both pieces of software are presented in Chapter 6.

Chapter 5 discusses the updating cycle in Bayesian statistics and focuses on the starting point of an updating cycle. The information in a Bayes factor is useful to describe how we can update our knowledge. However, knowing the rate with which the relative belief for two hypotheses changes is meaningless if the starting point is unknown. Chapter 5 therefore discusses the importance of prior probabilities and how to specify these for a set of hypotheses.

Chapters 7 and 8 present applied research where informative hypotheses are tested with Bayes factors. These are examples of research that commonly are analyzed with NHST and are thus exemplary in what the possibilities with informative hypothesis testing are. In Chapter 7 informative hypotheses are formulated to analyze the data from a repeated measures experiment. Chapter 8 evaluates the presence of a mediated effect at the the individual level by means of Bayesian informative hypothess tests.

Chapter 2

The power of informative hypotheses

by F. Klaassen, H. Hoijtink & X. Gu[1]

2.1 Introduction

Statistical analyses in behavioral research are often concerned with the comparisons between groups. For example, Monin, Sawyer, & Marquez (2008) were interested in the acceptance of moral rebels and conducted an experiment with four conditions. Half of the participants were asked to write and record a speech supporting a position they disagreed with (*actor condition*). After writing the speech, they were either shown a recording of an alleged previous participant that obeyed the task (*actor-obedient*) or of a moral rebel (*actor-rebel*) who refused to give the speech on the conflicting topic. The other half of the participants were given the instructions about writing and recording a speech allegedly given to other participants, but did not have to write a speech themselves (*observer condition*). After reading the instructions they too watched either an obedient previous 'participant' (*observer-obedient*) or a moral-rebel (*observer-rebel*). After watching the recording, participants rated how they perceived the person giving the speech.

A common approach is to analyze the resulting data with an ANOVA and test the null hypothesis that there is no difference between the four groups against the alternative hypothesis that there is *a* difference. This analysis does not evaluate any specific predictions based on theory, and the value of the conclusion of such a hypothesis test can be questioned (van de Schoot et al., 2011). A prediction can be translated into an informative hypothesis, that is, a hypothesis that describes the theoretical expectation of the researchers (Gu, Mulder, Deković, & Hoijtink, 2014; van de Schoot et al., 2011). For example, theory

[1]Manuscript under review at Psychonomic Bulletin & Review.
Author contributions: FK wrote the paper, R code and executed the simulations. XG and HH conceptualized the project, discussed progress and provided feedback on writing.

predicts an interaction between the role of the participant (observer/actor) and the role of the speaker (rebel/obedient) (Monin et al., 2008). Specifically, moral rebels are expected to be rejected by actors and appreciated by observers. An informative hypothesis in line with this theoretical expectation is

$$H : \mu_{\text{observer-rebel}} > \mu_{\text{actor-obedient}} > \mu_{\text{observer-obedient}} > \mu_{\text{actor-rebel}},$$

where μ is the average rated acceptance of the speaker in the corresponding condition. In this hypothesis the four group means are ordered from largest to smallest. A more general notation of this simple order constrained hypothesis (Kuiper & Hoijtink, 2010) is:

$$H_i : \mu_1 > ...\mu_k > ... > \mu_K, \tag{2.1}$$

where all K group means μ_k are ordered from large to small, with $k = 1, ..., K$. Throughout this paper, the focus is on hypotheses like H_i that describe an ordering of all K group means from large to small. Note that the concept of informative hypothesis is more general than constrain combinations of parameters by means of inequalities and equalities. However this is beyond the scope of this article.

Bayesian statistics can be used to find the best hypothesis from a set of competing hypotheses. The Bayes factor expresses the relative evidence in the data for two hypotheses (Kass & Raftery, 1995). Any informative hypothesis H_i can be compared to its complement H_c, to another hypothesis $H_{i'}$, where $i' \neq i$, or to the unconstrained hypothesis H_u (Hoijtink, 2012, pp. 50–51). The complement of H_i is H_c:

$$H_c : \text{ not } H_i, \tag{2.2}$$

which describes all other possible orderings of the parameters in H_i. A researcher can also compare H_i to another interesting hypothesis $H_{i'}$, any other ordering of the parameters, for example:

$$H_{i'} : \mu_2 > \mu_1 > ... > \mu_K. \tag{2.3}$$

Finally, a researcher can choose to compare H_i to H_u, the unconstrained hypothesis:

$$H_u : \mu_1, ..., \mu_k, ...\mu_K, \tag{2.4}$$

where all parameters can take on any value.

The Bayes factor BF_{ic} expresses the support in the data for H_i relative to H_{c_i}. For example, when $BF_{ic} = 5$, the support in the data for H_i is 5 times stronger than for H_{c_i}. When $BF_{ic} = 0.1$, the support for H_{c_i} is 10 times stronger than for H_i. A Bayes factor can be used to update prior odds into posterior odds. The prior odds is the ratio of the probability of H_i relative to the probability of $H_{i'}$ *before* observing the data. The posterior odds is the ratio of the probability of H_i relative to the probability of $H_{i'}$ *after* observing the data. Posterior probabilities are also referred to as *conditional error probabilities* (Berger, Boukai, & Wang, 1997; Hoijtink, 2012, pp. 80–81). If the posterior probabilities for H_i and H_c are .8 and .2, that is a posterior odds of 4, there is, given the data and prior probabilities, a probability of .8 that H_i is the best hypothesis and a probability of .2 that H_c is the best hypothesis. The *conditional* error probabilities depend on the chosen prior probabilities and the data

and provide meaningful information about the probability of a hypothesis *after* data are observed.

Unconditional error probabilities are well-known as the alpha-level and beta-level or the Type I and Type II error probabilities in the context of null-hypothesis significance testing. The unconditional error probabilities do not depend on the data and can be used to determine the required sample size to detect a particular effect size *prior* to observing data. Unconditional error probabilities are often used for sample size determination purposes in frequentist analyses, which is not common in Bayesian statistics. The focus in Bayesian hypothesis testing often lays in the conditional error probabilities. However, prior to data collection, unconditional error probabilities can provide information about what the expected strength of evidence is for a particular sample size. This paper will present methods that use both unconditional and conditional error probabilities to determine the sample size for the evaluation of informative hypotheses by means of Bayes factors.

Section 2.2 explains how the Bayes factor can be computed and used for comparing informative hypotheses. Section 2.3 presents an overview available sample size determination methods for Bayesian hypothesis testing. Different strategies are discussed that can be used to determine sample size based on unconditional or conditional error probabilities. In Section 2.4 four sample size determination approaches are introduced specifically for the comparison of informative hypotheses by means of Bayes factors. Three types of error (Type I, Type II and indecision error) and the desired level of evidence are combined. Section 2.5 describes how the methodology is programmed and the simulation conditions considered to illustrate the four approaches. The results of this simulation are presented and discussed in Section 2.6, followed by a set of guidelines for sample size determination in Bayesian informative hypothesis testing. Section 2.7 introduces three examples to illustrate these guidelines. Finally, Section 2.8 briefly discusses the findings of this paper.

2.2 Bayes factor

The Bayes factor is a tool for Bayesian hypothesis testing. Bayes factors can be computed for any pair of hypotheses, and can be used to quantify the evidence in favor of one of these hypotheses. Bayes factors penalize the fit with the complexity of the hypotheses under consideration, where the fit describes how well the data support a hypothesis, and the complexity describes how specific a hypothesis is. The Akaike Information Criterion (AIC) for example, is based on a similar principle. The AIC penalizes the maximum value of the likelihood, which is a measure of fit, by the number of parameters, which is a measure of complexity (Akaike, 1973). BF_{iu} can be expressed as a ratio of the fit f_i and the complexity c_i of H_i and expresses the support in the data for H_i relative to H_u (Hoijtink, 2012, pp. 51–52):

$$BF_{i1} = \frac{f_i}{c_i}. \tag{2.5}$$

Using BF_{iu} and BF_{cu} or $BF_{i'u}$, Bayes factors can be obtained that express the support in the data for H_i relative to H_c or $H_{i'}$:

$$BF_{ic} = \frac{BF_{i1}}{BF_{c1}} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}, \tag{2.6}$$

$$BF_{ii'} = \frac{BF_{i1}}{BF_{i'1}} = \frac{f_i}{c_i} / \frac{f_{i'}}{c_{i'}}. \tag{2.7}$$

In order to compute the fit and complexity of a hypothesis, the density of the data, and the prior and posterior distributions of the target parameters are needed. For an ANOVA model, the density of the data is:

$$f(\mathbf{y}|\boldsymbol{\mu}, \sigma^2) = \prod_{k=1}^{K} \prod_{s=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2} \frac{(y_{ks} - \mu_k)^2}{\sigma^2}), \tag{2.8}$$

where $\mathbf{y} = [y_{11}, ..., y_{1N}, ..., y_{K1}, ..., y_{KN}]$, $\boldsymbol{\mu} = [\mu_1, ..., \mu_K]$, σ^2 indicates the within group variance and is equal for each group, $k = 1, 2, \ldots, K$ indicates a group, and $s = 1, 2, \ldots, N$ indicates a person in group k. The sample size, denoted by N, is equal for each group.

Based on Gu et al. (2014) independent non-informative normal prior distributions are used for the parameters in the hypothesis, that is, the group means:

$$h(\boldsymbol{\mu}) = h(\mu_1) \cdot \cdots \cdot h(\mu_K), \tag{2.9}$$

with

$$h(\mu_k) = \mathcal{N}(0, \infty),$$

for $k = 1, ..., K$, in which the prior means are zero and the prior variances approach infinity.

The effect of this non-informative prior distribution on the posterior distribution is so small, that the posterior depends fully on the data. We use a normal approximation of the posterior distribution for the group means, that is, the target parameters:

$$g(\boldsymbol{\mu}|\mathbf{y}) = g(\mu_1) \cdot \cdots \cdot g(\mu_K), \tag{2.10}$$

with

$$g(\mu_k|\mathbf{y}) = \mathcal{N}(\hat{\mu}_k, \hat{\tau}_k^2),$$

for $k = 1, 2, ...K$, in which $\hat{\mu}_k$ is the estimate of the mean in group k, and $\hat{\tau}_k^2$ is the squared standard error of the mean in group k, where

$$\hat{\mu}_k = \frac{1}{N} \sum_{s=1}^{N} y_{ks}, \tag{2.11}$$

and

$$\hat{\tau}_k^2 = \frac{\sum_{s=1}^{N} (y_{ks} - \hat{\mu}_k)^2}{N \cdot (N - 1)}. \tag{2.12}$$

The complexity and fit of a hypothesis are based on the prior and posterior distribution.

The complexity of H_i, c_i, describes how specific H_i is. It is the proportion of the prior distribution in agreement with H_i (Hoijtink, 2012, p. 60):

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h(\boldsymbol{\mu}) d\boldsymbol{\mu}. \tag{2.13}$$

There are $K!$ unique hypotheses that order all parameters from small to large, each of which has the same complexity. Consequently, the complexity for hypotheses like H_i is $c_i = 1/K!$, and for H_c it is $c_c = 1 - c_i$ (Hoijtink, 2012, p. 60). Note that since H_c is the complement of H_i, $c_i + c_c = 1$.

The fit of H_i, f_i, describes how well the data support H_i. It is the proportion of the posterior distribution in agreement with H_i (Hoijtink, 2012, p. 59):

$$\begin{aligned} f_i &= \int_{\boldsymbol{\mu} \in H_i} g(\boldsymbol{\mu}|\mathbf{y}) d\boldsymbol{\mu} \\ &\approx \sum_{t=1}^{T} I_{\boldsymbol{\mu}_t \in H_i} / T, \end{aligned} \tag{2.14}$$

where $\boldsymbol{\mu}_t$ is sampled from $g(\boldsymbol{\mu}|\mathbf{y})$, $I_{\boldsymbol{\mu}_t \in H_i}$ is 1 if $\boldsymbol{\mu}_t$ is in agreement with H_i, and 0 otherwise, and T is the number of posterior samples. Again, since H_c is the complement of H_i, it follows that $f_c = 1 - f_i$. Using the complexity and fit, Bayes factors can be computed.

2.3 Sample size determination

The Bayes factor can be used to compute the conditional probabilities of the hypotheses under consideration. Often, the goal of hypothesis comparison is to not only describe the evidence in the data, but to select the best hypothesis from a set. If $BF_{12} = 1.1$ for example, this shows that the evidence is 1.1 times more in favor of H_1 relative to H_2. This corresponds to a conditional probability of approximately .52 for H_1 and .48 for H_2. These conditional error probabilities not provide any information about the the effect of the sample size on this conclusion. If the sample size in this example were 10, it seems very possible that the preference for H_1 is due to sampling variance. Alternatively, if the sample size were $10,000$, the preference for H_1 is more likely to be true in the population of interest. (Adcock, 1997) presents the first available research on the relation between sample size and the Bayes factor. Amongst others, he discusses the method of (Weiss, 1997).

Weiss (1997) advocates the importance of both conditional and unconditional power, and investigates different combinations of sample size, conditional and unconditional error probabilities. One of the approaches considers a cut-off of the Bayes factor such that the unconditional Type I error probability, that is, the probability that H_0 is preferred when H_u is true, is at the traditional .05. He creates sampling distributions for the Bayes factor for different sample sizes and true populations under H_u. From these sampling distributions he then derives the unconditional power. Using a cut-off for the Type I error probability determines a critical Bayes factor. Alternatively Weiss (1997) proposes to keep the cut-off of the Bayes factor fixed at 1, because this is a meaningful value, and determine the Type I and Type II error probabilities for this criterion. Not only does Weiss (1997) consider

both the conditional and unconditional error probabilities for different sample sizes, he presents multiple possible strategies for determining the sample size and discusses different populations to consider. This paper will elaborate on these different approaches. While they are only limited to the comparison of a null hypothesis to a one- or two sided alternative, this paper extends to the comparison of informative hypotheses.

De Santis (2004, 2007) present another Bayesian sample size determination on for the comparison of $H_0 : \mu = 0$ with $H_1 : \mu \neq 0$. This method applies a decision criterion where Bayes factors are only considered decisive if they are smaller than $\frac{1}{3}$ or larger than 3. The sample size is determined such that $P(BF_{01} > 3|H_0)$ and $P(BF_{01} < \frac{1}{3}|H_1)$ are both larger than a pre-specified value. In other words, an area of indecision is included in the determination of sample size that ensures that not both the unconditional and the conditional error probabilities are at a desired level. This strategy goes further than Weiss (1997), but is limited in two aspects. First, this approach does not include a limit on the unconditional probability that no decision is made. In other words, the sample size determination could potentially lead to a sample that gives a .05 Type I and Type II error probability, and an indecision probability of .9. In the current paper therefore, this approach is extended with the possibility to put a critical value on the indecision probability as well. Second, De Santis (2004, 2007) again only considers a single mean with a null and alternative hypothesis. Reyes & Ghosh (2013) consider Bayesian sample size determination methods for the difference between two means. One of their methods determines a critical Bayes factor such that the average error probability is minimized. The sample size is then determined such that average of the Type I and Type II error probability is smaller than a specified cut-off value. This idea will be incorporated in our proposed methods. The focus of these Bayesian sample size methods is on the null and alternative hypotheses.

Sample size determination for the evaluation of the null hypothesis H_0 with an informative hypothesis H_i using BF_{i0} is considered by Klugkist, Post, Haarhuis, & Wesel (2014). The decision criterion used is that Bayes factors larger and smaller than 1 result in conclusions in favor of H_i and H_0 respectively. Using this decision criterion, the sample size is determined for various effect sizes, such that the traditional Type I error probability is below .05, and the power is above .80 (Klugkist et al., 2014). Although this article uses order constrained hypotheses, no elaboration is made on the sample sizes required for the evaluation of H_i with H_c or with $H_{i'}$. Furthermore, the current research does not include a null hypothesis, so is focused on the sample size required for comparing informative hypotheses. The current research extends on this approach by considering not only the Type I and Type II error probability, but additionally the indecision and average error probabilities.

Other research discussing the relation between sample size and Bayes factors focuses on knowledge updating (e.g. Rouder, 2014). Specifically, this refers to the sequentially adding data and computing Bayes factors on this updated dataset to view how the evidence accumulates to the true hypothesis as more information is added. Schönbrodt & Wagenmakers (2018) extended this principle and investigated how large a sample should be to obtain 'strong' evidence, that is, to obtain a Bayes factor of a particular size. Multiple testing is a problem if sample size is determined for a desired level of unconditional error. However, if sample size is determined for a desired level of evidence there no longer is an effect of multiple testing and sequential analysis. Including the desired level of strength of evidence in the planning for sample size is relatively new to the literature on Bayesian sample size determination. The current research will include both strategies using

unconditional probabilities as a cut-off, and those using conditional error probabilities to determine sample size.

Concluding, all available existing strategies for sample size determination are limited the sense that they do not allow for the evaluation of two order constrained hypotheses. Hoijtink (2012, pp. 115–118) gives indications of appropriate sample sizes in this situation. Furthermore, the available papers mostly focus on one particular decision strategy. The variety in different approaches suggests that which error is the most detrimental to make can vary between situations. Therefore, we will elaborate on the available research by developing sample size tables for evaluation of H_i and H_c or H_i and $H_{i'}$ using different decision strategies. Particularly, we will develop four decision criteria for the Bayes factor, that incorporate both conditional and unconditional error probabilities.

2.4 Methods

All approaches make use of the sampling distributions of the Bayes factors under H_i and H_c, or under H_i and $H_{i'}$. Approach 1, like in Klugkist et al. (2014) and Weiss (1997), chooses H_i if $BF_{ic} > 1$ or $BF_{ii'} > 1$, and chooses H_c if $BF_{ic} < 1$ or $H_{i'}$ if $BF_{ii'} < 1$. Sample sizes will be determined such that the unconditional error probabilities are acceptably low. Approach 2, like in De Santis (2004) and De Santis (2007), chooses H_i if $BF_{ic} > 3$ or $BF_{ii'} > 3$, and chooses H_c if $BF_{ic} < \frac{1}{3}$ or $H_{i'}$ if $BF_{ii'} < \frac{1}{3}$. No decision is made if Bayes factors are between $\frac{1}{3}$ and 3. Again, sample sizes will be determined such that error probabilities are acceptably low. In Approach 3, the Bayes factor is not used to make a decision, but to express support for H_i and H_c or $H_{i'}$ based on the data. Sample sizes will be determined such that reasonably high Bayes factors can be expected, for example, 3, 10, or 20.

The sample size needed for the evaluation of H_i versus $H_{i'}$ or versus H_c can be determined such that error probabilities are acceptably low, or the median Bayes factor under the true hypothesis expresses acceptably strong support. This section will first explain how sampling distributions of Bayes factors are obtained. Second, each approach is explained in more detail, by precisely defining error probabilities and the median Bayes factor required. Finally, it will be described what is meant by acceptably low error probabilities and strong support. Throughout this section, the comparison of H_i and H_c using BF_{ic} is discussed. The discussion is analogous for H_i and $H_{i'}$, where all comments and notations regarding H_c can be replaced with corresponding ones regarding $H_{i'}$.

All approaches in this paper make use of the sampling distributions of Bayes factors. Amongst others, the effect sizes under H_i and under H_c need to be defined to obtain the sampling distributions. In this paper, Cohen's d, the standardized difference between two means, is used as a measure of effect size(Cohen, 1988, p. 276). The effect size d_{H_i} under H_i is the standardized difference between the largest and the smallest mean under H_i.

$$d_{H_i} = \frac{\mu_1 - \mu_K}{\sigma}, \tag{2.15}$$

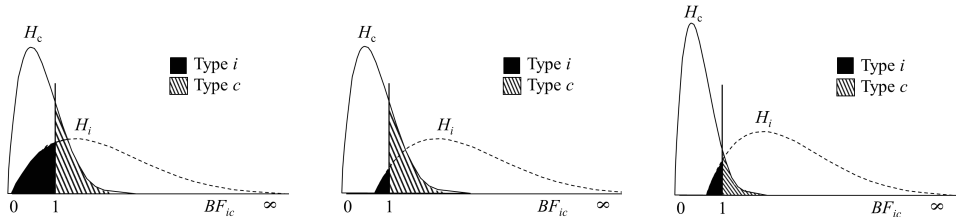
where μ_1 is the largest mean, and μ_K is the smallest mean under H_i. The effect size d_{H_c} under H_c is the standardized difference between the largest and the smallest population mean under H_c. For example, Figure 2.1 displays hypothetical sampling distributions of

BF_{ic} under H_i and under H_c, given $N = 50$, $d_{H_i} = .2$, and $d_{H_c} = .2$. These distributions represent the values of the Bayes factors observed if we repeatedly sample from populations under H_i and H_c. The procedure to obtain sampling distributions will be explained in full detail in Section 2.5.

2.4.1 Approach 1

The decision criterion used in Approach 1 is that H_i is preferred when BF_{ic} is larger than 1, and H_c is preferred when BF_{ic} is smaller than 1 (Klugkist et al., 2014; Weiss, 1997) . In Figure 2.1a, the vertical line at $BF_{ic} = 1$ indicates the decision criterion used in this approach: obtaining $BF_{ic} > 1$ results in the decision that the data support H_i, and $BF_{ic} < 1$ results in the decision that the data support H_c.

The vertical line marks two error probabilities. The first, the probability of observing $BF_{ic} < 1$ when H_i is true, $P(BF_{ic} < 1|H_i)$, is the probability of supporting H_c when H_i is true. In the remainder of this paper, this probability will be referred to as a Type i error probability. The second error probability is that of observing $BF_{ic} > 1$ when H_c is true denoted by $P(BF_{ic} > 1|H_c)$, that is, support for H_i when H_c is true. This will be referred to as Type c error probability. The average of Type i and Type c error probabilities will be called the *Decision error probability* which is similar to the average error probability used by Reyes & Ghosh (2013).



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

Figure 2.1. Error probabilities for Approach 1. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, given sample size N and effect sizes d_{H_i} and d_{H_c}. Note that $d_{H_c} = .2$ in each figure.

Figure 2.1b, if the effect size under H_i in Figure 2.1a increases, the sampling distribution under H_i shifts further away from the decision criterion, thus the Type i error decreases. As can be seen in Figure 2.1c, if the group sample size in Figure 2.1a increases, both Type i and Type c error decrease in this situation. For Approach 1, sample size will be determined such that the Type i, Type c, or Decision error probability is acceptably low.

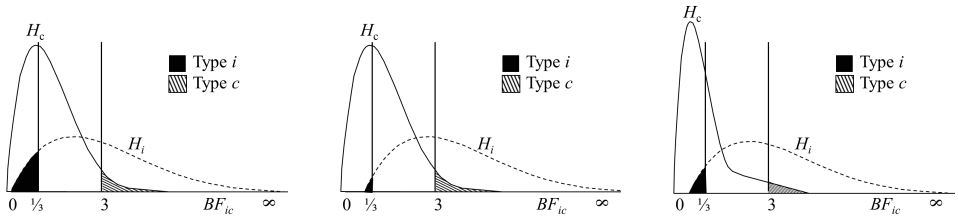
2.4.2 Approach 2

The decision criterion used in Approach 2 allows for indecision. Kass & Raftery (1995) have argued that Bayes factors between $\frac{1}{3}$ and 3 express too little support to prefer either

hypothesis. In Approach 2, like De Santis (2004) and De Santis (2007), this distinction is used by deciding that H_i is preferred for Bayes factors larger than 3 and deciding that H_c is preferred for Bayes factors smaller than $\frac{1}{3}$. For Approach 2, Type i error probability is expressed by $P(BF_{ic} < \frac{1}{3}|H_i)$ and Type c error probability by $P(BF_{ic} > 3|H_c)$. The average of Type i and Type c is the Decision error probability. An additional probability in this approach is that of not making a decision:

$$P(\tfrac{1}{3} < BF_{ic} < 3) = \frac{P(\frac{1}{3} < BF_{ic} < 3|H_i) + P(\frac{1}{3} < BF_{ic} < 3|H_c)}{2},$$

which is called the *Indecision probability*.



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

Figure 2.2. Error probabilities for Approach 2. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, for sample size N and effect sizes d_{H_i} and d_{H_c}. Note that $d_{H_c} = .2$ in each figure. The average of the area between $BF_{ic} = \frac{1}{3}$ and $BF_{ic} = 3$ under H_i and the area between $\frac{1}{3}$ and $BF_{ic} = 3$ under H_c, is the Indecision probability.

Figure 2.2 shows hypothetical sampling distributions of BF_{ic} under H_i and H_c and the error probabilities under Approach 2. As can be seen in Figure 2.2b, if the effect size under H_i in Figure 2.2a is increased, the Type i error probability decreases, while the Type c error probability remains constant. In Figure 2.2b it can also be seen that the Indecision probability decreases with the increased effect size. As can be seen in Figure 2.2c, if the sample size in Figure 2.2a is increased, the Type i and Type c error probabilities decrease. Since for both distributions, the size of the area between $\frac{1}{3}$ and 3 decreases, the Indecision probability also decreases. For Approach 2, sample size will be determined such that the Type i, Type c, or the Decision error probability is acceptably low. Based on the determined sample size and the decision criterion Indecision probability can be computed, but not controlled.

2.4.3 Approach 2b

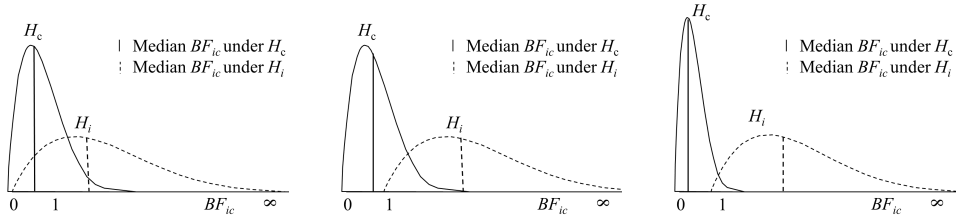
Note that the Indecision probability can be quite large in Approach 2, which might be undesirable for a researcher. Therefore, the situation in which a researcher wants to determine sample size such that the Indecision probability is acceptably low is also considered. We will refer to this approach by Approach 2b. In contrast to Approach 2, for Approach 2b sample size is determined such that the Indecision probability is controlled. Based on the sample size and decision criterion, the error probabilities can be determined,

but not controlled.

2.4.4 Approach 3

Approach 3 is different from Approach 1 and 2, because it does not rely on error probabilities or on a fixed decision criterion. In the sampling distributions under H_i and under H_c the median Bayes factor can be determined. These medians are an indication of the size of the Bayes factors that can be expected, given N, d_{H_i}, and d_{H_c}. The median was used, because it has the nice interpretation that exactly 50% of the distribution of Bayes factors is larger than the median, and 50% is smaller.

Figure 2.3 shows hypothetical sampling distributions of BF_{ic} under H_i and H_c. As can be seen in Figure 2.3a, each of the distributions is marked with a line, indicating the median value of that distribution. Note that in Approach 3, a researcher can choose a required value for the median Bayes factor under H_i or under H_c. As can be seen in Figure 2.3b, if the effect size in Figure 2.3a increases, the median Bayes factor under H_i increases, while the median Bayes factor under H_c remains constant. As can be seen in Figure 2.3c, if the sample size in Figure 2.3a increases, the median Bayes factor under H_i increases, while the median Bayes factor under H_c decreases. For Approach 3, sample size will be determined such that the median Bayes factor under H_i is of a required size, B, or the median Bayes factor under H_c is of a required size, $1/B$.



(a) $N = 50$ and $d_{H_i} = .2$. *(b)* $N = 50$ and $d_{H_i} = .5$. *(c)* $N = 100$ and $d_{H_i} = .2$.

Figure 2.3. Median Bayes factors for Approach 3. Hypothetical sampling distributions of BF_{ic} under H_i and H_c, given sample size N and effect size d_{H_i}. Note that $d_{H_c} = .2$ in each figure.

2.4.5 Critical values

Table 2.1 displays the critical values for the error probabilities, Indecision probability, and median Bayes factor considered in this paper. Note that traditionally in null hypothesis significance testing, Type I and Type II error probabilities are usually set at .05 and .2, resulting in an average error probability of .125. By limiting ourselves to Decision error probabilities of .1, .05, and .025 for Approach 1 and 2 (see Table 2.1), relatively strict cut-off values are used. We chose to do so, to respond to the replication crisis in social sciences. This crisis is partially due to publication of false positives (see for example Pashler & Wagenmakers (2012) and Thompson (2004)), which are partly caused by too

lenient Type I error rates. By using strict error probabilities, we determine sample sizes that have a relatively high probability of rendering correct results. For the Indecision probability in Approach 2b, .3, .2, and .1 are considered. Indecision probabilities larger than .3 have not been considered because then studies remain undecided too often. Furthermore, Indecision probabilities smaller than .1 were not considered, because then the Indecision probability becomes too small, and the situation resembles Approach 1 too much.

In Approach 3, 3, 10, and 20 are considered for B, roughly based on an indication of strength of support by Kass & Raftery (1995). A B of 3 implies a required median Bayes factor of 3 if H_i is true, and implies a required median Bayes factor of $1/B = 1/3$ if H_c is true. Note that a researcher could decide that both the Bayes factor if H_i is true and the Bayes factor if H_c is true, should be of a required size. This is done by determining the sample size such that the median Bayes factor under H_i is B, and the sample size such that the median Bayes factor under H_c is $1/B$. The largest of these two sample sizes is the required sample size.

Table 2.1
Critical error probabilities and critical B

Approach		Critical values		
1 and 2	Error probability	.1	.05	.025
2b	Indecision probability	.3	.2	.1
3	B	3	10	20

2.5 Simulation

Sample size tables are determined through simulations. The simulations are programmed and carried out in R (R Core Team, 2013). The hypotheses considered in this paper are H_i, H_c, and $H_{i'}$, like in Equations 2.1–2.2, with $K = 2, 3, 4$. The R code computes BF_{ic} or $BF_{ii'}$, based on samples from populations under H_i and H_c or under H_i and $H_{i'}$. The first three subsections describe in detail how the populations under H_i, H_c, and $H_{i'}$ are specified. These are the first steps of the simulation procedure. Section 2.5.4 gives a brief description of the entire simulation procedure by means of an example.

2.5.1 Specify H_i and effect size d_{H_i}

First, a population under H_i needs to be specified. The population is dependent on the number of groups under H_i, and on effect size d_{H_i}. As was indicated before, the effect size considered in this paper is Cohen's d. Based on Cohen's definition of small, medium, and large effect sizes, d_{H_i} can take on the values 0.2, 0.5, and 0.8 (Cohen, 1992). The group standard deviation σ_k is 1, for $k = 1, 2, ..., K$, and the smallest ordered mean is equal to 0. The difference between the first and the last ordered mean is described by d_{H_i}, and intermediate means are equally spaced between 0 and d_{H_i}. Table 2.2 shows the population

means for $K = 2, 3, 4$. If H_i is compared to H_c, $d_{H_i} = 0.2$, 0.5, and 0.8 are considered. If H_i is compared to $H_{i'}$, $d_{H_i} = 0.2$ and 0.5 are considered.

Note that because of our definition of effect size, the difference between each pair of means in a hypothesis for some effect size, varies over K. For example, for $K = 3$, and $d_{H_i} = .2$, the standardized difference between each pair of means is 0.1, while for $K = 4$, the difference is .067. We believe that by controlling the effect size over the difference between the first and the last mean, realistic mean orderings can be expressed. For example, for $K = 4$, it would be unrealistic to consider an effect size of 0.8 between each pair of means, because it would result in a standardized difference of 2.4 between the first and the last ordered mean. Although we believe our choices for effect size are realistic, we also acknowledge that we are being strict by considering rather small differences between pairs of means like .067.

Table 2.2
Population means given d

K	d	μ_1	μ_2	μ_3	μ_4
	0.2	0.2	0	-	-
2	0.5	0.5	0	-	-
	0.8	0.8	0	-	-
	0.2	0.2	0.1	0	-
3	0.5	0.5	0.25	0	-
	0.8	0.8	0.4	0	-
	0.2	0.2	0.133	0.067	0
4	0.5	0.5	0.333	0.167	0
	0.8	0.8	0.533	0.267	0

Note. d can be d_{H_i}, d_{H_c}, or $d_{H_{i'}}$. The means are labelled such that they match the ordering of means in H_i. The labels can be rearranged such that they match H_c or $H_{i'}$. For example, if $K = 3$, $d_{H_{i'}} = .2$, and $H_{i'} : \mu_3 > \mu_2 > \mu_1$, the populations means will be $\mu_3 = .2$, $\mu_2 = .1$, and $\mu_1 = 0$.

2.5.2 Specify H_c and effect size d_{H_c}

If H_i is evaluated with H_c, a population under H_c needs to be specified. The hypothesis H_c is the complement of H_i, indicating that every ordering of means not in H_i can be true. For $K = 2$, only one other ordering than that under H_i is possible, but five orderings are possible for $K = 3$, and 23 for $K = 4$. Table 2.3 shows all options of ordered means under H_c for $K = 2, 3$, and three examples for $K = 4$. As can be seen for $K = 3$, the orderings violate H_i in different ways. These violations are classified as small, medium, and large. An example of a small violation is a change in the order of only one pair of means, and an example of a large violation is a complete reversal of the ordering of means under H_i.

If a researcher is comparing H_i and H_c, he is testing an informative hypothesis H_i against its complement H_c, that is, he is testing one theory. The required sample size should be such that it can detect any deviation from his theory that is possible under H_c. Thus, additionally to a small effect size, researchers should always consider small violations under H_c. For a complete overview, this paper does present sample sizes required for medium and large

violations, too. Only $d_{H_c} = 0.2$ is considered. By doing so, the required sample sizes are sufficient to detect small deviations from H_i. We assume that if a researcher wants to evaluate H_i with H_c, he wants to be able to detect any deviation from his theory, specified in H_i. A small effect size under H_c renders a sample size sufficient to detect small deviations.

Table 2.3
Examples of ordered population means under H_c

K	Ordering	Violation of H_i
2	$\mu_2 > \mu_1$	-
3	$\mu_1 > \mu_3 > \mu_2$	small $*$
	$\mu_2 > \mu_1 > \mu_3$	small
	$\mu_2 > \mu_3 > \mu_1$	medium $*$
	$\mu_3 > \mu_1 > \mu_2$	medium
	$\mu_3 > \mu_2 > \mu_1$	large$*$
4	$\mu_1 > \mu_2 > \mu_4 > \mu_3$	small $*$
	$\mu_2 > \mu_3 > \mu_1 > \mu_4$	medium $*$
	$\mu_4 > \mu_3 > \mu_2 > \mu_1$	large $*$

Note. For $K = 4$ only a selection of ordered means is presented. An $*$ indicates that this ordering is used under $H_{i'}$. Note that $d_{H_c} = .2$, and $d_{H_{i'}} = .2$ or .5.

2.5.3 Specify $H_{i'}$ and effect size $d_{H_{i'}}$

If H_i is evaluated with $H_{i'}$, a population under $H_{i'}$ needs to be specified. To specify a population under $H_{i'}$, first a choice needs to be made for what ordering of means is considered under $H_{i'}$. Any ordering of means that is possible under H_c could be used as $H_{i'}$. In this paper, one ordering of means with a small violation of H_i is considered, one with a medium violation, and one with a large violation, for $K = 3, 4$. In Table 2.3 the orderings considered for $H_{i'}$ are marked with an asterisk.

If H_i is compared with $H_{i'}$, .2 and .5 are considered for both d_{H_i} and $d_{H_{i'}}$. We do so, because if a researcher wants to evaluate H_i with $H_{i'}$, he might value these two hypotheses equally. He can expect that a population under H_i is true, with for example an effect size of .5, but at the same time also consider a population under $H_{i'}$, with an effect size of .5.

2.5.4 Simulation procedure

This section describes the steps taken in the simulation procedure by means of an example. Figure 2.4 displays the simulation procedure, and highlights the choices made in the example.

1. Specify K, the number of groups, and the informative hypotheses considered: H_i, and H_c or $H_{i'}$. For this example, $K = 3$, $H_i : \mu_1 > \mu_2 > \mu_3$, which is compared with $H_c :$ not H_i.
2. Specify the effect sizes: d_{H_i} and d_{H_c} or $d_{H_{i'}}$. For this example, $d_{H_i} = .2$ and $d_{H_c} = .2$.

3. Determine the population means based on K, the effect sizes, and the hypotheses. As indicated before, throughout this paper, the group standard deviation is set to 1. If H_c is considered, the simulations have to be run in turn for each population mean ordering possible under H_c, using the same K, hypotheses, and effect sizes. This continues until all orderings have been considered. In the example, the second ordering from Table 2.3 is considered. The population means follow from Table 2.2.

4. Specify a starting group sample size N. In this example and for all simulations the starting group sample size is 2.

5. Sample J datasets using the population means and standard deviation, and sample size N. For all simulations, $J = 10,000$.

6. Compute the complexity and fit using Equation 2.13–2.14. Compute BF_{ic} or $BF_{ii'}$, using Equation 2.6 or 2.7. Since H_i is compared with H_c in this example, BF_{ic} is computed.

7. Compute the appropriate error probabilities, Indecision probability and the median Bayes factor for Approaches 1–3, based on the Bayes factors. Note that the Type i and Type c error probabilities are computed separately for Approach 1 and Approach 2, because of the different decision criteria (see Figure 2.1–2.3).

8. Increase the group sample size in Step 4 by 1, until $N = 1,000$, and repeat Steps 5–8.

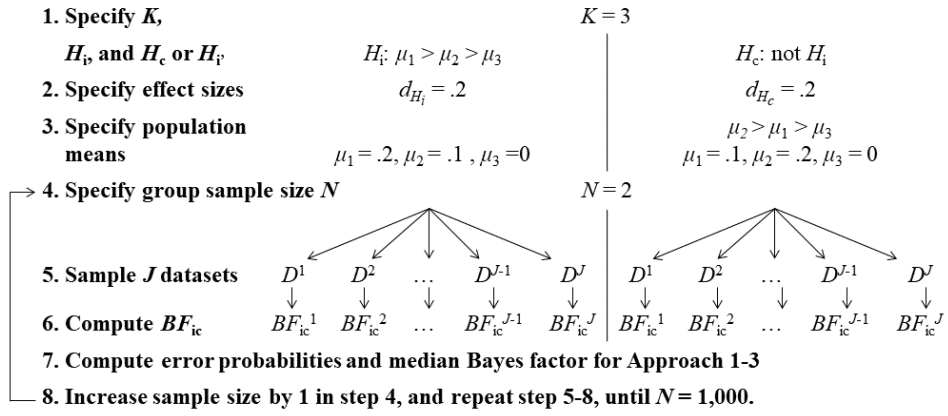


Figure 2.4. Example of the simulation procedure.

Based on the simulations the error probabilities and median Bayes factors for every group sample size from $N = 2$ to $N = 1,000$ are known. The required sample size can be determined based on the type and size of error one is willing to make (Approaches 1, 2, and 2b), or on the median Bayes factor (Approach 3). The critical error probabilities and median Bayes factors used are those presented in Table 2.1. If H_c is considered, the required sample size is determined for each of the orderings. Then, the orderings are grouped by violation size, and the average for each of these groups is computed. Thus, if two orderings exist with a small violation size, the average of the required sample sizes for these orderings is the required sample size for small violations.

2.6 Results

Tables 2.4–2.15 contain the required group sample sizes based on the simulations for each of the approaches. Separate tables are presented for the comparison of H_i and H_c, and for the comparison of H_i and $H_{i'}$. The tables and the examples can contain very small required sample sizes. Note that it is advised to use a group sample size of at least 10, even if a table or an example suggests a smaller sample size. We provide a minimum, because inferences based on small sample sizes are susceptible to outliers. Sections 2.6.1–2.6.4 illustrate by means of brief examples how each table can be used. Section 2.6.5 interprets these tables. Note that because Sections 2.6.1–2.6.4 are to some extent repetitive, it may very well be that you want to skip to Section 2.6.5.

2.6.1 Approach 1

Tables 2.4 and 2.5 show the required group sample sizes using Approach 1, with $K = 2, 3, 4$, for the evaluation of H_i and H_c with BF_{ic}, and with $K = 3, 4$, for the evaluation of H_i and $H_{i'}$ with $BF_{ii'}$.

Example 1.1 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Decision error probability at .05. He specifies $d_{H_i} = .5$, $d_{H_c} = .2$, and expects even small violations to be possible under H_c. As can be seen in Table 2.4, the required sample size is 977. Suppose this researcher did not consider H_c, but $H_{i'}$ with a small deviation of H_i. As can be seen in Table 2.5, the required sample size is 327.

Example 1.2 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Type c error probability at .1. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$, and expects that only large violations are possible under H_c. As can be seen in Table 2.4, the required sample size is 54. Suppose this researcher did not consider H_c, but $H_{i'}$ with a large deviation of H_i. As can be seen in Table 2.5, the required sample size is 81.

Example 1.3 Suppose a researcher wants to evaluate H_i with H_c, with $K = 4$. The researcher wants to control the Type i error probability at .025. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$. As can be seen in Table 2.4, the required sample size is 443. Suppose this researcher did not consider H_c, but $H_{i'}$. As can be seen in Table 2.5, the required sample size is larger than $1,000$ if $H_{i'}$ was specified with a small violation of H_i, the sample size is 575 with a medium violation, and 169 with a large violation.

Table 2.4

Required group sample sizes for Approach 1 using H_c

		Error probability											
		.1			.05			.025			.1	.05	.025
K	d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8	-	-	-
2		82	40	36	132	83	82	187	132	132	82	132	187
		83	14	6	132	23	9	187	32	12			
3	*s*	644	624	624	994	**977**	977	*	*	*	977	*	*
	m	136	56	43	222	110	103	313	180	180	103	180	255
	l	109	35	24	181	65	58	278	108	100	**54**	100	148
		159	28	12	258	42	17	361	58	24			
4	*s*	*	*	*	*	*	*	*	*	*	*	*	*
	m	353	274	271	547	486	486	749	690	690	486	690	*
	l	146	47	31	241	92	78	345	151	142	76	141	206
		207	35	15	314	51	21	**443**	75	30			

Note. Non shaded areas give the sample size needed to control the Decision error probability, gives the sample size needed to control the Type i error probability, and gives the sample size needed to control the Type c error probability. Let * denote group sample sizes larger than 1,000. Small, medium, and large violations under H_c are denoted by s, m, and l. Note that s gives the average over the required group sample sizes for all population mean orderings under H_c that are a small violation of H_i. This is analogous for medium and large violations. Note that $d_{H_c} = .2$ for all sample sizes.

2.6.2 Approach 2

Tables 2.6 and 2.7 show the required group sample sizes using Approach 2, with $K = 2, 3, 4$, for the evaluation of H_i and H_c with BF_{ic}, and with $K = 3, 4$, for the evaluation of H_i and $H_{i'}$ with $BF_{ii'}$. Tables 2.8 and 2.9 present the corresponding Indecision probabilities.

Example 2.1 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Decision error probability at .05. He specifies $d_{H_i} = .5$, $d_{H_c} = .2$, and expects even small violations to be possible under H_c. As can be seen in Tables 2.6 and 2.8, the required sample size is 451, and the Indecision probability is .2. Suppose this researcher did not consider H_c, but $H_{i'}$ with a small deviation of H_i. As can be seen in Table 2.7 and 2.9, the required sample size is 64, and the Indecision probability is .363.

Example 2.2 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Type c error probability at .1. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$, and expects that only large violations are possible under H_c. As can be seen in Table 2.6 and 2.8, the required sample size is 3, and the Indecision probability is .435. Suppose this researcher did not consider H_c, but $H_{i'}$ with a large deviation of H_i. As can be seen in Table 2.7 and 2.9, the required sample size is 34, and the Indecision probability is .256.

Example 2.3 Suppose a researcher wants to evaluate H_i with H_c, with $K = 4$. The researcher wants to control the Type i error probability at .025. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$. As can be seen in Table 2.6 and 2.8, the required sample size is 233, and the Indecision probability is .160. Suppose this researcher did not consider H_c, but $H_{i'}$. As can be seen in Table 2.7 and 2.9, the required sample size is 628 if $H_{i'}$ was specified with a small violation of H_i with an Indecision probability of .429, and the required sample size is 318 with a medium violation with an Indecision probability of .265, and a sample size of 114 with a large violation with an Indecision probability of .144.

Table 2.5

Required group sample sizes for Approach 1 using $H_{i'}$

				Error probability								
				.1		.05		.025		.1	.05	.025
K		$d_{H_{i'}}$	d_{H_i} =	.2	.5	.2	.5	.2	.5	-	-	-
		.2		318	147	531	**327**	731	531	318	531	731
	s	.5		147	51	327	88	531	117	51	88	117
		-		318	51	531	88	731	117			
		.2		103	54	180	108	252	180	108	180	249
3	m	.5		54	18	103	29	180	40	17	30	41
		-		103	18	180	29	252	40			
		.2		81	41	135	81	192	135	**81**	135	192
	l	.5		41	14	81	22	135	31	14	22	31
		-		81	14	135	22	192	31			
		.2		725	338	*	725	*	*	725	*	*
	s	.5		338	115	725	189	*	273	115	189	273
		-		725	115	*	189	*	273			
		.2		247	111	399	233	562	382	228	382	540
4	m	.5		124	40	257	63	415	91	38	64	88
		-		257	41	415	63	**575**	91			
		.2		73	37	124	73	169	124	73	124	169
	l	.5		36	13	73	20	123	30	13	20	30
		-		73	13	123	20	**169**	28			

Note. Non shaded areas give the sample size needed to control the Decision error probability, gives the sample size needed to control the Type i error probability, and gives the sample size needed to control the Type i' error probability. Let * denote group sample sizes larger than $1,000$. Small, medium, and large violations under $H_{i'}$ are denoted by s, m, and l.

Table 2.6

Required group sample sizes for Approach 2 using H_c

		Error probability											
		.1			.05			.025			.1	.05	.025
groups	d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8		-	
2		20	8	6	48	23	20	77	50	48	20	48	77
		20	4	2	48	8	4	77	15	6			
3	*s*	175	39	10	459	**451**	451	730	730	730	451	730	985
	m	30	9	5	72	22	14	130	48	36	6	34	78
	l	22	7	4	58	17	10	103	30	19	**3**	14	32
		51	9	5	99	18	8	169	27	12			
4	*s*	*	*	*	*	*	*	*	*	*	*	*	*
	m	94	19	8	233	154	146	388	332	329	140	329	494
	l	44	11	6	96	24	13	163	46	30	2	23	61
		80	16	7	146	24	12	**233**	38	17			

Note. Non shaded areas give the sample size needed to control the Decision error probability, gives the sample size needed to control the Type *i* error probability, and gives the sample size needed to control the Type *c* error probability. Let * denote group sample sizes larger than $1,000$. Small, medium, and large violations under H_c are denoted by *s*, *m*, and *l*. Sample size is determined for each ordering of population means under H_c, and then averaged and presented in violation categories. Note that $d_{H_c} = .2$ for all sample sizes.

Table 2.7

Required group sample sizes for Approach 2 using $H_{i'}$

				Error probability								
				.1		.05		.025		.1	.05	.025
K		$d_{H_{i'}}$	d_{H_i} =	.2	.5	.2	.5	.2	.5	-	-	-
		.2		46	17	147	**64**	291	147	46	147	291
	s	.5		17	9	64	29	147	49	9	29	49
		-		46	9	147	29	291	49			
		.2		42	19	86	46	147	88	42	86	143
3	m	.5		18	8	45	16	86	24	8	15	25
		-		42	8	86	17	147	24			
		.2		34	16	72	35	115	72	**34**	72	115
	l	.5		16	7	35	13	72	19	7	13	19
		-		34	7	72	13	115	19			
		.2		69	26	313	133	628	327	69	313	628
	s	.5		26	15	133	51	327	104	15	51	104
		-		69	15	313	51	**628**	104			
		.2		76	30	176	87	314	174	68	167	294
4	m	.5		31	14	91	30	189	51	14	30	49
		-		80	14	189	33	**318**	52			
		.2		40	19	76	43	114	76	40	76	114
	l	.5		19	8	43	13	76	20	8	13	20
		-		40	8	76	14	**114**	20			

Note. Non shaded areas give the sample size needed to control the Decision error probability, gives the sample size needed to control the Type i error probability, and gives the sample size needed to control the Type c error probability. Let * denote group sample sizes larger than $1,000$. Small, medium, and large violations under $H_{i'}$ are denoted by s, m, and l.

Table 2.8

Indecision probabilities for Approach 2 using H_c

		Error probability											
		.1			.05			.025			.1	.05	.025
K	d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8		-	
2		.422	.384	.343	.328	.282	.226	.262	.181	.164	.226	.164	.131
		.422	.423	.377	.328	.384	.367	.262	.332	.343			
3	*s*	.389	.375	.416	.236	**.200**	.200	.162	.149	.149	.200	.149	.109
	m	.477	.472	.437	.388	.386	.344	.286	.268	.225	.435	.226	.159
	l	.455	.459	.441	.371	.385	.363	.271	.307	.271	**.435**	.316	.200
		.380	.449	.421	.292	.388	.395	.187	.330	.338			
4	*s*	*	*	*	*	*	*	*	*	*	*	*	*
	m	.427	.421	.400	.297	.242	.226	.199	.158	.152	.259	.152	.103
	l	.395	.407	.389	.311	.340	.317	.219	.241	.219	.340	.269	.155
		.340	.379	.381	.240	.337	.325	**.160**	.273	.280			

Note. Non shaded areas give the Indecision probability belonging to the sample size needed to control the Decision error probability, gives the Indecision probability belonging to the sample size needed to control the Type i error probability, and gives the Indecision probability belonging to the sample size needed to control the Type c error probability. Let * denote group sample sizes larger than $1,000$. Small, medium, and large violations under H_c are denoted by s, m, and l. Indecision probability is determined based on the sample size for each ordering of population means under H_c, and then averaged and presented in violation categories. Note that $d_{H_c} = .2$ for all sample sizes.

Table 2.9

Indecision probability for Approach 2 using $H_{i'}$

				Error probability								
				.1		.05		.025		.1	.05	.025
K		$d_{H_{i'}}$	d_{H_i} =	.2	.5	.2	.5	.2	.5	-	-	-
		.2		.519	.506	.401	**.363**	.289	.233	.531	.467	.403
	s	.5		.506	.496	.363	.381	.233	.281	.496	.381	.281
		-		.531	.496	.467	.381	.403	.281			
3	m	.2		.279	.250	.204	.159	.128	.104	.308	.251	.222
3	m	.5		.252	.264	.161	.184	.106	.129	.264	.169	.129
		-		.304	.264	.265	.189	.221	.123			
		.2		.242	.211	.174	.139	.115	.089	**.256**	.222	.199
	l	.5		.211	.214	.139	.149	.089	.109	.214	.149	.109
		-		.256	.214	.222	.149	.199	.109			
		.2		.593	.566	.428	.390	.298	.245	.592	.512	.429
	s	.5		.566	.557	.389	.420	.245	.293	.557	.420	.293
		-		.592	.557	.512	.420	**.429**	.293			
		.2		.352	.323	.238	.200	.147	.123	.378	.317	.257
4	m	.5		.318	.324	.194	.233	.121	.150	.324	.218	.146
		-		.375	.324	.316	.233	**.265**	.154			
		.2		.183	.147	.120	.090	.080	.061	.187	.168	.144
	l	.5		.147	.152	.090	.111	.061	.071	.152	.111	.071
		-		.188	.152	.168	.105	**.144**	.071			

Note. Non shaded areas give the Indecision probability belonging to the sample size needed to control the Decision error probability, gives the Indecision probability belonging to the sample size needed to control the Type i error probability, and gives the Indecision probability belonging to the sample size needed to control the Type c error probability. Let * denote group sample sizes larger than $1,000$. Small, medium, and large violations under $H_{i'}$ are denoted by s, m, and l.

Table 2.10
Required group sample sizes for Approach 2b using H_c

		Indecision probability								
		.3			.2			.1		
K	d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8
2		60	19	9	108	44	30	182	108	108
	s	304	74	34	567	**427**	427	*	*	*
3	*m*	119	40	20	194	72	46	325	154	141
	l	90	31	16	158	55	32	**281**	102	83
	s	294	67	29	*	366	353	*	*	*
4	*m*	218	**55**	23	367	177	154	622	492	491
	l	102	32	15	181	59	33	324	124	101

Note. Let * denote group sample sizes larger than 1,000. Small, medium, and large violations under H_c are denoted by *s*, *m*, and *l*. Sample size is determined based on the allowed Indecision probability for each ordering of population means under H_c, and then averaged and presented in violation categories. Note that $d_{H_c} = .2$ for all sample sizes.

2.6.3 Approach 2b

Tables 2.10 and 2.11 show the required group sample sizes using Approach 2b, with $K = 2, 3, 4$, for the evaluation of H_i and H_c with BF_{ic}, and with $K = 3, 4$, for the evaluation of H_i and $H_{i'}$ with $BF_{ii'}$. Tables 2.12 and 2.13 depict the corresponding Type *i*, Type *c* or Type *i*′, and Decision error probability.

Example 2b.1 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Indecision probability at .2. He specifies $d_{H_i} = .5$, $d_{H_c} = .2$, and expects even small violations to be possible under H_c. As can be seen in Table 2.10 and 2.12, the required sample size is 427, and the Type *i* error probability is smaller than .001, Type *c* is .112, and the Decision error probability is .056. Suppose this researcher did not consider H_c, but $H_{i'}$ with a small deviation of H_i. As can be seen in Table 2.11 and 2.13, the required sample size is 190, and the Type *i* error probability is .001, Type *c* is .044, and the Decision error probability is .023.

Example 2b.2 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to control the Indecision probability at .1. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$, and expects that only large violations are possible under H_c. As can be seen in Table 2.10 and 2.12, the required sample size is 281, and the Type *i* error probability is .008, Type *c* is smaller than .001, and the Decision error probability is .004. Suppose this researcher did not consider H_c, but $H_{i'}$ with a large deviation of H_i. As can be seen in Table 2.11 and 2.13, the required sample size is 61, and the Type *i* error probability is .059, Type *c* is smaller than .001, and the Decision error probability is .030.

Example 2b.3 Suppose a researcher wants to evaluate H_i with H_c, with $K = 4$. The researcher wants to control the Indecision probability at .3. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$, and expects medium violations of H_i under H_c. As can be seen in Table 2.10 and 2.12, the required sample size is 55, and the Type *i* error probability is .012, Type *c* is .118, and the Decision error probability is .065. Suppose this researcher did not consider H_c, but $H_{i'}$ with a medium violation of H_i. As can be seen in Table 2.11 and 2.13, the required sample size is 2 and the Type *i* error probability is .231, Type *c* is .383, and the Decision error probability is .307.

Table 2.11

Required group sample sizes for Approach 2b using $H_{i'}$

				Indecision probability					
				.3		.2		.1	
K		$d_{H_{i'}}$	d_{H_i} =	.2	.5	.2	.5	.2	.5
3	s	.2		266	97	442	**190**	746	447
		.5		97	44	190	71	447	118
	m	.2		2	2	87	31	180	92
		.5		2	2	31	14	91	30
	l	.2		2	2	55	19	127	61
		.5		2	2	19	9	**61**	21
4	s	.2		600	222	990	443	*	990
		.5		222	100	443	160	990	270
	m	.2		2	**2**	223	86	428	227
		.5		2	2	87	37	239	68
	l	.2		2	2	2	2	93	38
		.5		2	2	2	2	38	16

Note. Let * denote group sample sizes larger than 1,000. Small, medium, and large violations under $H_{i'}$ are denoted by s, m, and l.

Table 2.12

Error probabilities for Approach 2b using H_c

				Indecision probability								
				.3			.2			.1		
K			d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8
2		i		.038	.014	.008	.014	.001	.000	.005	.000	.000
		c		.038	.102	.132	.014	.051	.075	.005	.014	.014
		DE		.038	.058	.070	.014	.026	.037	.005	.007	.007
3	s	i		.008	.003	.002	.001	**.000**	.000	*	*	*
		c		.142	.201	.192	.080	**.112**	.112	*	*	*
		DE		.075	.102	.097	.041	**.056**	.056	*	*	*
	m	i		.042	.010	.008	.021	.003	.000	.006	.000	.000
		c		.016	.045	.066	.006	.027	.041	.001	.010	.012
		DE		.029	.028	.037	.014	.015	.021	.004	.005	.006
	l	i		.056	.020	.013	.028	.005	.002	**.008**	.001	.000
		c		.009	.027	.047	.002	.018	.024	**.000**	.005	.008
		DE		.033	.024	.030	.015	.012	.013	**.004**	.003	.004
4	s	i		.015	.007	.006	*	.000	.000	*	*	*
		c		.364	.273	.212	*	.364	.365	*	*	*
		DE		.190	.140	.109	*	.182	.182	*	*	*
	m	i		.031	**.012**	.011	.012	.001	.000	.003	.000	.000
		c		.076	**.118**	.118	.044	.087	.093	.015	.026	.026
		DE		.053	**.065**	.065	.028	.044	.047	.009	.013	.013
	l	i		.082	.035	.028	.039	.010	.004	.012	.002	.000
		c		.014	.038	.055	.005	.024	.039	.001	.010	.014
		DE		.048	.037	.042	.022	.017	.022	.007	.006	.007

Note. Type i, Type c, and Decision error probability are denoted by i, c, and DE. Small, medium, and large violations under H_c are denoted by s, m, and l. The error probabilities are determined based on the required sample size for each ordering of population means under H_c, and then averaged and presented in violation categories. Note that $d_{H_c} = .2$ for all sample sizes. Let * indicate that sample sizes larger than 1,000 were required to meet this level of Indecision probability, and thus no error probabilities are known.

Table 2.13
Error probabilities for Approach 2b using $H_{i'}$

				Indecision probability					
				.3		.2		.1	
K		$d_{H_{i'}}$ =	d_{H_i} =	.2	.5	.2	.5	.2	.5
3	s	.2	i	.030	.007	.015	**.001**	.005	.000
			i'	.030	.068	.015	**.044**	.005	.014
			DE	.030	.037	.015	**.023**	.005	.007
		.5	i	.068	.028	.044	.015	.014	.005
			i'	.007	.028	.001	.015	.000	.005
			DE	.037	.028	.022	.015	.007	.005
	m	.2	i	.289	.214	.053	.016	.018	.046
			i'	.322	.322	.053	.121	.018	.000
			DE	.305	.268	.053	.069	.018	.023
		.5	i	.289	.214	.122	.054	.046	.016
			i'	.239	.239	.014	.058	.001	.016
			DE	.264	.227	.068	.056	.024	.016
	l	.2	i	.316	.226	.070	.022	.022	.000
			i'	.328	.328	.070	.146	.022	.059
			DE	.322	.277	.070	.084	.022	.030
		.5	i	.316	.226	.146	.074	**.059**	.020
			i'	.239	.239	.022	.074	**.000**	.020
			DE	.277	.232	.084	.074	**.030**	.020
4	s	.2	i	.030	.007	.014	.001	*	.000
			i'	.030	.064	.014	.038	*	.014
			DE	.030	.035	.014	.020	*	.007
		.5	i	.064	.027	.038	.014	.014	.005
			i'	.007	.027	.001	.014	.000	.005
			DE	.035	.027	.019	.014	.007	.005
	m	.2	i	.255	**.222**	.043	.009	.016	.000
			i'	.410	**.410**	.038	.094	.013	.040
			DE	.333	**.316**	.040	.052	.015	.020
		.5	i	.255	.222	.095	.044	.040	.016
			i'	.346	.346	.009	.038	.000	.014
			DE	.300	.284	.052	.041	.020	.015
	l	.2	i	.320	.231	.320	.231	.037	.004
			i'	.383	.383	.383	.383	.038	.107
			DE	.352	.307	.352	.307	.038	.056
		.5	i	.320	.231	.320	.231	.107	.036
			i'	.291	.291	.291	.291	.004	.037
			DE	.306	.261	.306	.261	.056	.037

Note. Type i, Type c, and Decision error probability are denoted by i, c, and DE. Small, medium, and large violations under $H_{i'}$ are denoted by s, m, and l. Let * indicate that sample sizes larger than $1,000$ were required to meet this level of Indecision probability, and thus no error probabilities are known.

2.6.4 Approach 3

Table 2.14–2.17 show the required group sample sizes for $K = 2, 3, 4$, for the evaluation of H_i and H_c with BF_{ic}, using Approach 3. Tables 2.14 and 2.15 show the required group sample sizes if the median BF_{ic} or $BF_{ii'}$ is required to be of size B under H_i, and Tables 2.16 and 2.17 show the required group sample sizes if the median BF_{ic} or $BF_{ii'}$ is required to be of size $1/B$ under H_c or $H_{i'}$.

Example 3.1 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to know that 50% of the possible Bayes factors if H_i is true, is larger than 10, and that 50% of the possible Bayes factors if H_c is true, is smaller than $\frac{1}{10}$. He specifies $d_{H_i} = .5$, $d_{H_c} = .2$, and expects even small violations to be possible under H_c. As can be seen in Tables 2.14 and 2.16, the required sample sizes to meet the boundaries are 50 and 839, respectively. Because the researchers wants to adhere to both boundaries, the largest sample size is required, which is 839. Additionally, we know that 13.4% of the possible Bayes factors under H_c is larger than 1, which implies that the probability of finding evidence in favour of the H_i when H_c is true, is .134.

Suppose this researcher did not consider H_c, but $H_{i'}$ with a small deviation of H_i. As can be seen in Table 2.15 and 2.17, the required samples size are 63 and 391, respectively. Again, because both boundaries need to be adhered, the largest sample size must be considered, which is 391. Additionally, we know that 7.8% of the possible Bayes factors under H_c is larger than 1, thus a probability of .078 to find evidence in favour of H_i when H_c is true.

Example 3.2 Suppose a researcher wants to evaluate H_i with H_c, with $K = 3$. The researcher wants to know that 50% of the possible Bayes factors if H_c is true, is smaller than $\frac{1}{3}$. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$, and expects that only large violations are possible under H_c. As can be seen in Table 2.16, the required sample size is 2, and additionally, we know that 30.3% of possible Bayes factors under H_c is larger than 1. This researcher has a probability of .303 to find evidence in favour of H_i when H_c is true. Suppose this researcher did not consider H_c, but $H_{i'}$ with a large deviation of H_i. As can be seen in Table 2.15, the required sample size is 9, and additionally, we know that 33.2% of possible Bayes factors under $H_{i'}$ is larger than 1. This researcher has a probability of .332 to find evidence in favour of H_i when H_c is true.

Example 3.3 Suppose a researcher wants to evaluate H_i with H_c, with $K = 4$. The researcher wants to know that 50% of possible Bayes factors if H_i is true, is larger than 20. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$. As can be seen in Table 2.14 the required sample size is 637, and additionally, we know that 1.1% of all possible Bayes factors under H_i is smaller than 1. Suppose this researcher did not consider H_c, but $H_{i'}$. As can be seen in Table 2.15 the required sample size is 38 using a large violation of H_i under $H_{i'}$, with 18% of possible Bayes factors under H_i smaller than 1. For medium violations, the sample size is 283, with 9.1% of possible Bayes factors smaller than 1, and for small violations, the sample size required is larger than $1,000$.

2.6.5 Discussion of table features

This section discusses two features of the tables presented in Section 2.6. First, the required sample sizes are compared to sample sizes presented by Cohen (1992) for the evaluation of

Table 2.14

Required group sample sizes for controlling BF_{ic} under H_i at B in Approach 3

		B								
		3			10			20		
K	d_{H_i} =	.2	.5	.8	.2	.5	.8	.2	.5	.8
2		24	4	2	87	15	6	136	23	9
		.242	*.238*	*.210*	*.091*	*.089*	*.091*	*.044*	*.046*	*.038*
3		81	13	6	301	**50**	20	513	83	33
		.196	*.210*	*.184*	*.039*	***.037***	*.037*	*.011*	*.010*	*.012*
4		97	16	7	352	58	23	**637**	103	40
		.220	*.224*	*.231*	*.040*	*.039*	*.040*	***.011***	*.010*	*.010*

Note. Entries in italics indicate $P(BF_{ic} < 1|H_i)$. Note that $d_{H_c} = .2$ for all sample sizes.

Table 2.15

Required group sample sizes for controlling $BF_{ii'}$ under H_i at B in Approach 3

			B					
			3		10		20	
K		d_{H_i}	.2	.5	.2	.5	.2	.5
3	*s*		118	19	391	**63**	577	94
			.219	*.221*	*.078*	***.079***	*.043*	*.040*
	m		17	2	67	10	107	17
			.305	*.331*	*.159*	*.173*	*.103*	*.107*
	l		9	2	41	6	67	11
			.332	*.332*	*.183*	*.195*	*.123*	*.126*
4	*s*		274	45	884	144	✻	212
			.218	*.213*	*.079*	*.072*	–	*.040*
	m		56	2	187	30	**283**	45
			.280	*.345*	*.137*	*.138*	***.091***	*.092*
	l		2	2	24	2	**38**	2
			.394	*.296*	*.230*	*.296*	***.180***	*.296*

Note. Entries in italics indicate $P(BF_{ic} < 1|H_i)$. Let * denote sample sizes larger than $1,000$, and – denote the absence of $P(BF_{ic} < 1|H_i)$ in this situation.

Table 2.16

Required group sample sizes for controlling BF_{ic} under H_c at $1/B$ in Approach 3

		B		
K		3	10	20
2		24	87	136
		.242	*.091*	*.044*
3	*s*	444	**839**	*
		.281	***.134***	–
	m	21	110	165
		.247	*.095*	*.061*
	l	**2**	58	101
		.303	*.091*	*.049*
4	*s*	2	*	*
		.303	–	–
	m	2	370	492
		.272	*.146*	*.099*
	l	2	52	107
		.228	*.146*	*.071*

Note. Entries in italics indicate $P(BF_{ic} > 1|H_c)$. Let * denote sample sizes larger than $1,000$, and – denote the absence of $P(BF_{ic} > 1|H_c)$ in this situation. Note that $d_{H_c} = .2$ for all sample sizes.

Table 2.17

Required group sample sizes for controlling $BF_{ii'}$ under $H_{i'}$ at $1/B$ in Approach 3

K			3	10	20
3	*s*	.2	118	**391**	577
			.219	***.078***	*.043*
		.5	19	63	94
			.221	*.079*	*.040*
	m	.2	17	67	105
			.303	*.158*	*.103*
		.5	3	11	17
			.312	*.161*	*.100*
	l	.2	**9**	41	67
			.332	*.183*	*.123*
		.5	2	6	11
			.334	*.196*	*.126*
4	*s*	.2	274	884	*
			.218	*.079*	–
		.5	45	144	212
			.214	*.073*	*.040*
	m	.2	50	179	271
			.282	*.131*	*.085*
		.5	9	30	43
			.287	*.125*	*.082*
	l	.2	4	24	40
			.385	*.231*	*.175*
		.5	2	4	6
			.359	*.242*	*.196*

Note. Entries in italics indicate $P(BF_{ic} > 1|H_c)$. Let * denote sample sizes larger than $1,000$, and – denote the absence of $P(BF_{ic} > 1|H_c)$ in this situation.

Table 2.18
Required sample sizes

	Critical value		Effect size		
Approach	Type	Size	.2	.5	.8
Cohen (1992)	Decision error	.125	393	64	26
Approach 1	Decision error	.100	82	40	36
Approach 2	Decision error	.100	20	8	6
	Indecision		*.422*	*.384*	*.343*
Approach 2b	Indecision	.100	182	108	108
	Decision error		*.005*	*.007*	*.007*
Approach 3	B under H_i	10	87	15	6
	$P(BF_{ic} < 1\|H_i)$		*.091*	*.089*	*.091*
Approach 3	$1/B$ under H_c	$\frac{1}{10}$	87		
	$P(BF_{ic} > 1\|H_c)$		*.091*		

Note. Effect size indicates d for Cohen (1992), and d_{H_i} for Approach 1–3. Note that $d_{H_c} = .2$ for all approaches. Note that Cohen's approach compares H_0 and H_1, while Approaches 1–3 compare H_i and H_c. Entries in italics are additional probabilities rendered by an approach.

H_0 and H_1. Secondly, the benefit of using $H_{i'}$ over H_c is discussed.

The sample sizes presented in Tables 2.4–2.17 might seem large on first view. However, in this paper strict measures for the effect sizes and the error probabilities have been used. Small, medium, and large effect sizes are used, however, these effect sizes describe the difference between the largest and the smallest mean. Thus, large differences between each pair of means are not common. As was explained in Section 2.4.5, the used critical values in this paper (.1, .05 and .025) are more strict than the Decision error probability based on the traditional Type I and Type II error probabilities ((.05 + .2)/2 = .25/2 = .125).

In order to put the results obtained in this paper in perspective, the sample sizes based on the approaches in this paper are compared with the sample sizes presented by Cohen (1992). Table 2.18 presents required sample sizes for $K = 2$, using Cohen's comparison of H_0 and H_1, and for each of the approaches presented in this paper. Specifically, all approaches are compared to the sample sizes for the evaluation of $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$, with a Type I error probability of .05, and a Type II error probability of .80, that is, a Decision error probability of .125. Only $K = 2$ is considered, because for $K = 2$, the effect sizes used in this paper corresponds to the effect sizes used by Cohen. Furthermore, although the comparison of $H_i : \mu_1 > \mu_2$ with $H_c : \mu_1 < \mu_2$ is different from the comparison of H_0 with H_1, it does give an impression of how the required sample sizes compare.

For each approach, the type and size of the critical value is specified such that the method is as similar as possible to the Decision error probability used by Cohen. A Decision error probability under Approach 1 is most comparable to that of Cohen when it is .100,

since larger values than .100 are not considered in this paper. A Decision error probability under Approach 2 is most comparable to that of Cohen when it is .100. Approach 2b seemed most comparable to Cohen when an Indecision probability of .100 is considered, since additionally to the Indecision probability, this approach renders Decision error probability. For Decision error probabilities smaller than .025, Approach 2b results in a higher probability of a correct decision than Cohen's setup. Finally, in Approach 3 B does not correspond to a Decision error of a certain size. Therefore, $B = 10$ is considered, which is usually considered to express a fair amount of evidence.

As can be seen in Table 2.18, for a small effect size under H_i and H_c, the required sample size comparing H_i and H_c with Approaches 1–3 is smaller than required for comparing H_0 and H_1 using Cohen's method. Furthermore, for all approaches but Approach 2b, the required sample size for a medium effect size is smaller than that required for Cohen's approach.

The comparison with the sample sizes prescribed for null hypothesis significance testing puts the results of this paper in perspective. For small to medium effect sizes, which are often expected in applied research, Approaches 1–3 require smaller sample sizes than Cohen's power analysis. For Approach 2, it appears that the smaller sample sizes do come at the cost of a relatively large Indecision probability. For large effect sizes, the required sample sizes are smaller for Approach 2 and 3 relative to Cohen's sample size, and for the other approaches, the sample sizes are not much larger than those following from Cohen (1992). For $K = 3, 4$, the sample sizes are less easy to compare, because of different uses of effect size, and more complex hypotheses. However, if the results are compared, comparing H_i to H_c will require similar sample sizes to Cohen, whereas comparing H_i to $H_{i'}$ will in general result in smaller sample sizes.

As can be seen in the tables, in general, a smaller sample size is required if H_i is compared to $H_{i'}$ than when it is compared to H_c. For example, as can be seen in Table 2.4, the required sample size for $K = 3$, $d_{H_i} = .5$, and a Decision error probability of .05, the required sample size is 977 for small violations of H_c, which is the only violation that should be considered in practice. As can be seen in Table 2.5, if $H_{i'}$ is considered, the sample size ranges from 22 to 327, dependent on the choice of violation size and effect size under $H_{i'}$. All of these sample sizes are much smaller than the 977 required for the comparison of H_i to H_c. Thus, if you have a competing theory, you are better off using $H_{i'}$ than H_c. Note that the required sample size is not in all situations smaller when using $H_{i'}$ rather than H_c. Appendix 10.1 explains situations in which this is not the case, and further elaborates on some numerical characteristics of the tables.

2.7 In practice

This section provides guidelines for applied researchers to select an approach, $H_{i'}$ or H_c, an effect size, and a critical value. Figure 2.5 shows a decision tree, with some example research questions. First of all, the decision tree will be discussed, and then the further choices that must be made.

As can be seen in Figure 2.5, the choice for an approach depends on maximally two sequential questions. The first question, *What type of decision do you want to make?* relates

to whether a dichotomous, trichotomous, or no decision should be made. For dichotomous decisions, that is, choosing between H_i and H_c or $H_{i'}$, Approach 1 applies. For trichotomous decisions, that is, choosing between H_i, H_c or $H_{i'}$, and indecision, either Approach 2 or 2b applies. For situations in which a researcher does not want to make decision, but express the support in the data for each hypothesis, Approach 3 applies. If a trichotomous decision is required, the second question, *What probability do you want to control for?* has to be answered. This relates to whether a researcher wants to control the Indecision probability, that is, Approach 2b, or control the Type i, c, i' or Decision error probability, that is, Approach 2.

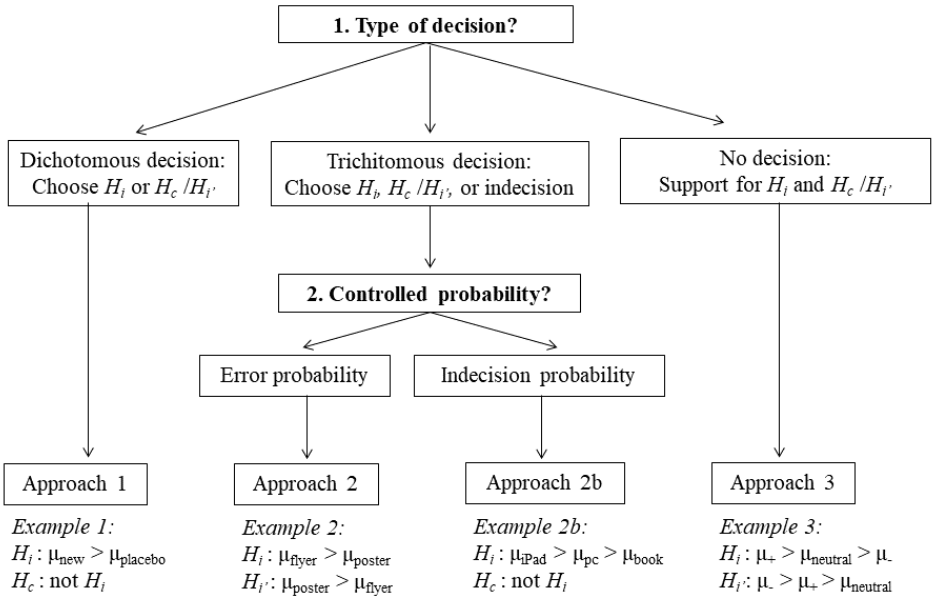


Figure 2.5. Decision Tree

Example 1. Suppose a researcher wants to see if a new drug is more effective than a placebo, $H_i : \mu_{\text{new}} > \mu_{\text{placebo}}$, and compares this with the complement, H_c. It is very important to know if H_i or H_c is true, to support the decision to implement the drug or not. Answering Question 1 in Figure 2.5 this researcher would need to use Approach 1 to determine the required group sample size, because a dichotomous decision has to be made. (cf. Tables 2.4–2.5).

Example 2. Suppose a researcher wants to investigate whether flyers or posters are more effective in informing inhabitants of a neighbourhood about upcoming events, $H_i : \mu_{\text{flyer}} > \mu_{\text{poster}}$ versus $H_{i'} : \mu_{\text{poster}} > \mu_{\text{flyer}}$. The researcher wants to make a decision for H_i or Hi' only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Finally, he does not have a limit to what indecision he maximally allows, so he replies to Question 2 that he wants to control the error probability. This researcher would need to use Approach 2 to determine the required group sample size (cf. Tables 2.6–2.9).

Example 2b. Suppose a researcher wants to investigate the effect of learning tool on the test outcome of students. He hypothesizes $H_i : \mu_{\text{iPad}} > \mu_{\text{PC}} > \mu_{\text{book}}$, and H_c : not H_i. The researcher wants to make a decision for H_i or H_c only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Because his research is quite costly to execute, he wants to limit the Indecision probability. Therefore, this researcher should use Approach 2b to determine the required group sample size (cf. Tables 2.10–2.13).

Example 3. Suppose a researcher wants to evaluate two competing theories. The theories concern the attitude of people towards healthy food, after being primed with positive, neutral, or negative cues. He hypothesizes $H_i : \mu_+ > \mu_{\text{neutral}} > \mu_-$ and $H_{i'} : \mu_- > \mu_+ > \mu_{\text{neutral}}$. This researcher is not interested in making a decision, but wants to express the support in the data for H_i and $H_{i'}$. Following Question 1 in Figure 2.5, he needs to use Approach 3 to determine the required sample size (cf. Tables 2.16–2.17).

After determining the appropriate approach, a researcher still needs to make three decisions. First of all, a researcher needs to decide whether he wants to compare H_i to H_c or $H_{i'}$. If H_c is used, as explained in Section 2.5.2, only small violations of H_i should be considered, and if $H_{i'}$ is used, the researcher must decide based on his theory, what the ordering of means under $H_{i'}$ is. Table 2.3 displays what is considered a small violation under H_c, and shows the orderings considered under $H_{i'}$ in this paper

Secondly, a researcher needs to choose the effect sizes and population means under H_i and H_c or $H_{i'}$. Table 2.2 displays the population means for the effect sizes considered in this paper. Inspiration for effect size can be taken from previous research in the same field. If the effect size generally is .5, use .5. If no previous research exists, it is up to the researcher to choose a reasonable effect size. It is advised to use a small effect size in this situation.

Thirdly, a researcher needs to make one or two decisions regarding the critical value. This differs per approach. Table 2.1 displays the critical values for the different decision criteria used in this paper. For Approach 1 and 2, a researcher must first decide whether he wants to control Type i, Type c or Type i', or Decision error probability. This choice is dependent on what type of error the researcher values more strongly. For example, if a Type i error is deemed most harmful, the Type i error probability must be controlled. Secondly, the researcher must choose the critical value. This should be done based on practical value. The smaller the value, the larger the probability that the resulting decision will be correct.

For Approach 2b, a researcher must only decide what critical value he considers for the Indecision probability. This choice depends on the costs related to not making a decision. If the costs are high, a small critical value should be chosen for the Indecision probability.

For Approach 3, a researcher must first decide whether he wants to control the median Bayes factor under H_i, the median Bayes factor under H_c or $H_{i'}$, or control both. For example, if the evidence under H_i is deemed most important, the chosen B only refers to Bayes factors under H_i. Secondly, the researcher must choose a size of this median Bayes factor, which is expressed by B. This should be done based on practical value. Tentative guidelines for the strength of the evidence expressed by B can be found in Kass & Raftery (1995). According to them, $B = 3$ expresses positive support, and $B = 20$ expresses strong support.

2.8 Discussion

In this paper, sample sizes have been determined for the comparison of H_i with H_c or $H_{i'}$, by means of three main approaches. As was indicated in Section 2.3 and 2.4, strict effect sizes and critical values have been used. The effect sizes have been chosen such that they gave a reasonable representation of what can be expected in social sciences. For the error probabilities, it should be noted again that strict error probabilities are required for more sound research outcomes. In order to accommodate researchers that want different effect sizes, error probabilities, or different orderings of means under $H_{i'}$ than those considered in this paper, an R script has been developed[2].

Future research could investigate the required sample sizes for different types of hypotheses. For example, the use of a composite $H_{i'}$, that consists of multiple orderings, can be of interest for researchers that consider not all orderings under H_c relevant, but do not have one specific theory. Furthermore, other informative hypotheses than simple order constrained hypotheses could be considered. Finally, only have been considered ANOVA models. It would be interesting to extend this research to other statistical models.

[2]The script and the manual can be downloaded using this link: 10.17605/OSF.IO/D9EAJ

Chapter 3

All for one or some for all? Evaluating informative hypotheses using multiple $N = 1$ studies

by F. Klaassen, C. Zedelius, H. Veling, H. Aarts and H. Hoijtink[1]

3.1 Introduction

There is increasing attention for individual centered analyses (e.g. Molenaar, 2004; Hamaker, 2012). For example, in personalized medicine it is not relevant to find if a treatment works *on average* in a group of individuals but rather whether it works for any individual (Woodcock, 2007). This paper is concerned with individual centered analyses in the form of multiple $N = 1$ studies. A core feature of this paper is that multiple hypotheses are formulated for each person. These hypotheses are first evaluated at the individual level and subsequently conclusions are formed at the group level. Specifically, this will be done in the context of a within-subject experiment (for a pilot study into using informative hypothesis in the context of multiple $N = 1$ studies, see Kluytmans et al., n.d.). In a within-subject experiment each person $i = 1, ..., P$ is exposed to the same set of experimental conditions $j = 1, ..., J$. By conducting R replications with a dichotomous outcome (0 = failure, 1 = success) in

[1]Published as Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H., & Hoijtink, H. (2017). All for one or some for all? Evaluating informatie hypotheses using multiple N = 1 studies. *Behavior Research Methods, 50*(6), pp. 2276-2291.

Author contributions: FK wrote the paper, designed, programmed, and executed the simulation. HH provided input for the project. HH and FK further conceptualized this project. HH provided feedback on writing, concepts, and programming. CZ, HV and HA provided the data and hypotheses used in the illustration and feedback on the writing.

condition j the number of successes x_j^i of person i can be obtained. This can be modeled using a binomial model with R trials and unknown success probability π_j^i.

This paper proposes a Bayesian method that evaluates informative hypotheses (Hoijtink, 2012) for multiple within-subject $N = 1$ studies. Researchers can formulate informative hypotheses based on (competing) theories or expectations. This can be achieved by using the relations '>' and '<' to impose constraints on the parameters $\boldsymbol{\pi}^i = [\pi_1^i, ..., \pi_J^i]$. E.g. '$\pi_1^i > \pi_2^i$' states that π_1^i is larger than π_2^i and reversely, '$\pi_1^i < \pi_2^i$' states that π_1^i is smaller then π_2^i. When a comma is used to separate two parameters, such as 'π_1^i, π_2^i', no constraint is imposed between these parameters. For each person, multiple informative hypotheses can be evaluated by means of Bayes factors (Kass & Raftery, 1995). Using the Bayes factor, it can be determined for each person which hypothesis is most supported by the data. Here our method departs from traditional analyses. Rather than evaluating hypotheses at the group level, the hypotheses are evaluated for each person separately. In social psychology, for example, it is often hoped or thought that if a hypothesis holds at the group level, this also applies to all individuals (see for example, Moreland & Zajonc, 1982; Klimecki, Mayer, Jusyte, Scheeff, & Schönenberg, 2016). Hamaker (2012) describes the importance of individual analyses using an example: Cross-sectionally, the number of words typed per minute and the percentage of typos might be negatively correlated. That is, people that type fast tend to be good at typing and thus make fewer mistakes than people that type slow. However at the individual level a positive correlation exists between these variables, *i.e.* if a fast typer goes faster than his normal typing speed, the number of mistakes will increase (Hamaker, 2012). Similarly, if multiple persons aim to score a penalty several times, we might find that the average success probability is smaller than .5, however this does not imply that each individual has a penalty scoring probability smaller than .5. Differently from Hamaker (2012) and Molenaar (2004) our approach does not stop at a single $N = 1$ study. Rather, when individual analyses have been executed, it is interesting to see if all individuals support the same hypothesis. Thus, when multiple hypotheses are evaluated for P individuals, two types of conclusions can be drawn. First, by executing multiple $N = 1$ studies it can be determined for each person if any hypothesis can be selected as the best, and if so, which hypothesis this is. Second, it can be determined if the sample comes from a population that is homogeneous with respect to the support of the specified hypotheses, and if so, which hypothesis is supported most.

This paper is structured as follows. First, the difference between analyses at the group level and multiple $N = 1$ analyses is elaborated upon by means of an example that will be used throughout the paper. Second, it will be described how informative hypotheses can be evaluated for one $N = 1$ study. Third, it will be explained how multiple $N = 1$ studies can be used to evaluate each hypothesis and detect if any can be selected as the best hypothesis for all individuals. The appropriate number of replications and the number of participants can be determined using a sensitivity analysis. The paper is concluded with a short discussion.

3.2 P-population and WP-population

An example of a within-subject experiment is Zedelius, Veling, & Aarts (2011). These researchers investigated the effect of interfering information and reward on memory. In

each trial, participants were shown five words on a screen and asked to remember these for a brief period of time. During this time interfering information was presented on the screen. Afterwards they were asked to recall the five words verbally in order to obtain a reward. Three factors with two levels each were manipulated over the trials: Before each trial started, participants were shown a *high* (hr) or a *low* (lr) reward on the screen they would receive upon completing the task correctly. This reward could be displayed *subliminally* (sub), that is, very briefly (17ms) or *supraliminally* (sup), that is for a longer duration of 300ms. Finally, the visual stimulus interfering with the memory task was either a sequence of letters, *low interference* (li), or eight words that were different from the five memorized *high interference* (hi). Combining these factors results in eight conditions, for example *hr-sub-hi* and *lr-sup-li*. Seven trials were conducted in each condition, resulting in a total of 56 trials per participant. After each trial the participant was given a score of 1 if all five words were recalled and 0 if not.

Zedelius et al. (2011) specified expectations regarding the ordering of success probabilities that can be translated in many different hypotheses. One example of an informative hypothesis based on the expectations of Zedelius et al. (2011) is

$$\begin{aligned} H_1 :& \textit{hr-sup-li} > \textit{hr-sup-hi} > \textit{hr-sub-li} > \textit{hr-sub-hi} > \\ & \textit{lr-sup-li} > \textit{lr-sup-hi} > \textit{lr-sub-li} > \textit{lr-sub-hi}, \end{aligned} \tag{3.1}$$

where *hr-sup-li* is $\pi_{\text{hr-sup-li}}$, the success probability in condition hr-sup-li. For simplifications in the remainder of this paper, π is omitted in the notation of all examples using the conditions from Zedelius et al. (2011). Alternatively, for each person i the hypothesis could be formulated as:

$$\begin{aligned} H_1^i :& \textit{hr-sup-li}^i > \textit{hr-sup-hi}^i > \textit{hr-sub-li}^i > \textit{hr-sub-hi}^i > \\ & \textit{lr-sup-li}^i > \textit{lr-sup-hi}^i > \textit{lr-sub-li}^i > \textit{lr-sub-hi}^i, \end{aligned} \tag{3.2}$$

where $\textit{hr-sup-li}^i$ is the success probability in condition hr-sup-li of person i.

To illustrate the difference between Equation 3.1 and 3.2 let us consider a *population of persons* (P-population from hereon) and a *within-person population* (WP-population from hereon). Each individual in the P-population has their own success probabilities $\boldsymbol{\pi}^i$. The averages of these individual probabilities are the P-population probabilities $\boldsymbol{\pi} = [\pi_1, ..., \pi_J]$, where $\pi_j = \frac{1}{P}\sum_{i=1}^{P} \pi_j^i$. Equation 3.1 is a hypothesis regarding the ordering of these P-population probabilities. Equation 3.2 is a hypothesis regarding the ordering of the WP-population probabilities for person i. Evaluating this hypothesis for person i is an example of an $N = 1$ study.

Many statistical methods are suited to draw conclusions at the P-population level. However, if a hypothesis is true at the P-population level, there is no guarantee that it holds for all WP-populations (Hamaker, 2012). Thus, a conclusion at the P-population level does not necessarily apply to each individual. Rather than $\boldsymbol{\pi}$, this paper concerns the individual $\boldsymbol{\pi}^i$. If multiple hypotheses are formulated for each person i, it can be determined for each person which hypothesis is most supported. Furthermore, it can be assessed whether the sample of P persons comes from a population that is homogeneous with respect to the informative hypotheses under consideration.

3.3 N = 1: How to analyze the data of one person

This section describes how the data of one person can be analyzed. First, the general form of hypotheses considered for every person are introduced. Subsequently, the statistical model used to model the $N = 1$ data is introduced. Finally, the Bayes factor is introduced and elaborated upon.

3.3.1 Hypotheses

Researchers can formulate informative hypotheses regarding $\boldsymbol{\pi}^i$. The general form of the informative hypotheses used in this paper is:

$$H_m^i : R_m \boldsymbol{\pi}^i > 0, \tag{3.3}$$

where $m, m' = 1, ..., M (m \neq m')$ is the label of a hypothesis, M is the number of hypotheses considered and m' is another hypothesis than m, $\boldsymbol{\pi}^i = [\pi_1^i, ...\pi_J^i]$ and R_m is the constraint matrix with J columns and K rows, where K is the number of constraints in a hypothesis. The constraint matrix can be used to impose constraints on (sets of) parameters. An example of a constraint matrix R for $J = 4$ is:

$$R_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \tag{3.4}$$

which renders

$$H_1^i : \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i, \tag{3.5}$$

which specifies that the success probabilities $\boldsymbol{\pi}^i$ are ordered from large to small. Note that the first row of R_1 specifies that $1 \cdot \pi_1^i - 1 \cdot \pi_2^i + 0 \cdot \pi_3^i + 0 \cdot \pi_4^i > 0$, that is, $\pi_1^i > \pi_2^i$. The constraint matrix

$$R_2 = \begin{bmatrix} .5 & .5 & -.5 & -.5 \end{bmatrix}, \tag{3.6}$$

renders the informative hypothesis

$$H_2^i : \frac{\pi_1^i + \pi_2^i}{2} > \frac{\pi_3^i + \pi_4^i}{2}, \tag{3.7}$$

which states that the average of the first two success probabilities is larger than the average of the last two. Hypotheses constructed using Equation 3.3 are a translation of the expectations researchers have with respect to the outcomes of their experiment into restrictions on the elements of $\boldsymbol{\pi}^i$.

Another hypothesis that is considered in this paper is the complement of an informative hypothesis:

$$H_{\not m}^i : \text{not } H_m^i. \tag{3.8}$$

The complement states that H_m^i is not true in the WP-population. Stated otherwise, the reverse of the researchers' expectation is true. Finally, H_u^i denotes the unconstrained

hypothesis:

$$H_u^i : \pi_1^i, \pi_2^i, ..., \pi_{J-1}^i, \pi_J^i, \tag{3.9}$$

where each parameter is 'free'. An informative hypothesis H_m^i constrains the parameter space such that only particular combinations of parameters are allowed, $H_{m'}^i$ comprises that part of the parameter space that is not included in H_m^i and the conjunction of H_m^i and $H_{m'}^i$ is H_u^i. The difference in use of H_u^i and $H_{m'}^i$ will be elaborated further in the section on Bayes factors.

Zedelius et al. (2011) formulated several expectations concerning the ordering of success probabilities over the experimental conditions. The main expectation was that high reward trials would have a higher success probability than low reward trials. This main effect and the expectations regarding the other conditions (interference level and visibility duration) can be translated in various informative hypotheses (Kluytmans et al., n.d.). A first translation of the expectations is

$$\begin{aligned} H_1^i : &\textit{hr-sup-li}^i > \textit{hr-sup-hi}^i > \textit{hr-sub-li}^i > \textit{hr-sub-hi}^i > \\ &\textit{lr-sup-li}^i > \textit{lr-sup-hi}^i > \textit{lr-sub-li}^i > \textit{lr-sub-hi}^i, \end{aligned} \tag{3.10}$$

which states that for any person i the success probabilities are ordered from high to low. To give some intuition for this hypothesis, Figure 3.1 shows eight bars that represent the experimental conditions, and its height indicates the success probability in that condition, and the ordering of probabilities adheres to H_1^i. Substantively, this hypothesis specifies that all conditions with a high reward have a higher success probability than those with a low reward, which in Figure 3.1 can be verified since all dark gray bars are higher than any light gray bar. Furthermore, H_1^i specifies that within this main reward value effect, that is, looking only at high reward success conditions or only at low reward conditions, a supraliminally shown rewards (solid border) results in a higher success probability than a subliminally shown reward (dotted border). Finally, within the visibility duration effect, that is, looking only at conditions with the same reward and same visibility duration, low interference (no pattern) results in a higher success probability than high interference (diagonally striped pattern). Alternatively, two less specific hypotheses can be formulated that include the main effect of reward and only one of the remaining main effects:

$$H_2^i : \textit{hr-li}^i > \textit{hr-hi}^i > \textit{lr-li}^i > \textit{lr-hi}^i, \tag{3.11}$$

and

$$H_3^i : \textit{hr-sup}^i > \textit{hr-sub}^i > \textit{lr-sup}^i > \textit{lr-sub}^i, \tag{3.12}$$

where $\textit{hr-li}^i$ indicates the average success probability of the $\textit{hr-sup-li}^i$ and $\textit{hr-sub-li}^i$ conditions. In Figure 3.1, both H_2^i and H_3^i are true. Different from H_1^i, these hypotheses do not state that *any* high reward condition has a higher success probability than *any* low reward condition, but rather that averaged over both interference level and visibility duration high reward conditions have a higher success probability than low reward conditions. Additionally, H_2^i further specifies that averaged over visibility duration, the success probability is always higher in high reward conditions compared to low reward conditions. Within this main effect of reward value the success probability is higher for

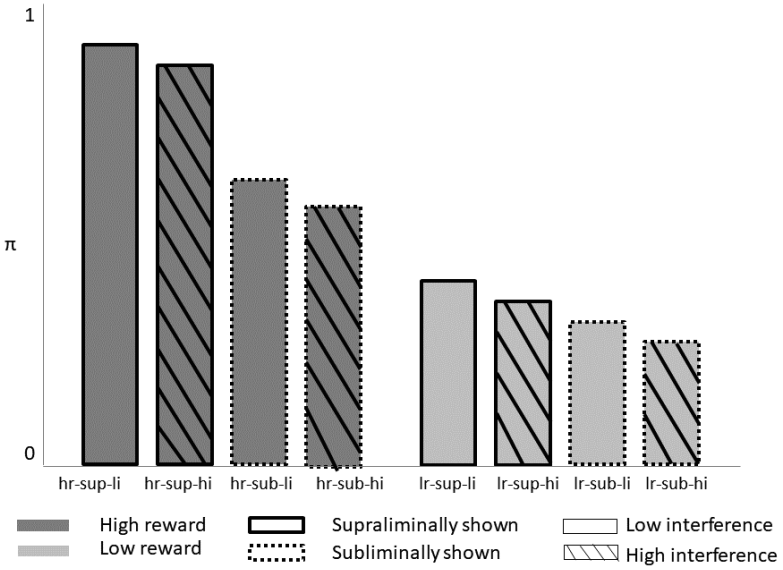


Figure 3.1. Graphical representation of all hypotheses by Zedelius et al. (2011)

low interference than for high interference. Analogously, H_3^i states that averaged over interference level, the success probability is always larger in high compared to low reward conditions. Within this pattern the success probability is larger for supraliminally compared to subliminally shown rewards.

A fourth hypothesis relates to the interaction effect between reward type and visibility duration:

$$H_4^i : hr\text{-}sup^i - lr\text{-}sup^i > hr\text{-}sub^i - lr\text{-}sub^i, \tag{3.13}$$

which states that the benefit of high reward over low reward is larger when the reward is shown supraliminally compared to when the reward is shown subliminally. This, too, is presented in Figure 3.1, since the difference between *hr-sup* (average of the dark-gray, solid border bars) and *lr-sup* (average of the light-gray, solid border bars) is larger than the difference between *hr-sub* (average of the dark-gray, dashed border bars) and *lr-sub* (average of the light-gray, dashed border bars). Note that, other than H_2^i and H_3^i, H_1^i is not a special case of H_4^i. These hypotheses can both be true, as is presented in the figure, but knowing that H_1^i is true gives no information about H_4^i.

Together H_1^i, H_2^i, H_3^i and H_4^i form a set of competing informative hypotheses that can be evaluated for each person.

3.3.2 Density, prior, posterior

To evaluate hypotheses using a Bayes factor, the density of the data, prior and posterior distribution are needed. For the type of data used in this paper, that is, the number of successes $\mathbf{x}^i = [x_1^i, ..., x_J^i]$ observed for person i in R replications in each condition j the density of the data is

$$f(\mathbf{x}^i \mid \boldsymbol{\pi}^i) = \prod_{j=1}^{J} \binom{R}{x_j^i} (\pi_j^i)^{x_j^i} (1 - \pi_j^i)^{R - x_j^i}, \tag{3.14}$$

that is, in each condition j the response x_j^i is modeled by a binomial distribution. The prior distribution $h(\boldsymbol{\pi}^i | H_u^i)$ for person i is a product over Beta distributions

$$h(\boldsymbol{\pi}^i \mid H_u^i) = \prod_{j=1}^{J} \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} (\pi_j^i)^{\alpha_0 - 1} (1 - \pi_j^i)^{\beta_0 - 1}, \tag{3.15}$$

where $\alpha_0 = \beta_0 = 1$, such that $h(\boldsymbol{\pi}^i \mid H_u^i) = 1$, that is, a uniform distribution. As will be elaborated upon in the next section, only $h(\boldsymbol{\pi}^i \mid H_u^i)$ is needed for the computation of the Bayes factors involving H_m^i, $H_{m'}^i$ and H_u^i (Klugkist et al., 2005). The interpretation of α_0 and β_0 is the prior number of successes and failures plus one. In other words, using $\alpha_0 = \beta_0 = 1$ implies that the prior distribution is uninformative. Consequently, the posterior distribution based on this prior is completely determined by the data. Furthermore, by using $\alpha_0 = \beta_0 = 1$ for each $\boldsymbol{\pi}^i$ the prior distribution is unbiased with respect to informative hypotheses that belong to an equivalent set (Hoijtink, 2012, p. 205). As will be elaborated in the next section, unbiased prior distributions are required to obtain Bayes factors that are unbiased with respect to the informative hypotheses under consideration.

The unconstrained posterior distribution is proportional to the product of the prior distribution and the density of the data:

$$\begin{aligned} g(\boldsymbol{\pi}^i \mid \mathbf{x}^i, H_u^i) &\propto f(\mathbf{x}^i \mid \boldsymbol{\pi}^i) \cdot h(\boldsymbol{\pi}^i \mid H_u^i) \\ &\propto \prod_{j=1}^{J} \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} (\pi_j^i)^{\alpha_1 - 1} (1 - \pi_j^i)^{\beta_1 - 1}, \end{aligned} \tag{3.16}$$

where $\alpha_1 = x_j^i + \alpha_0 = x_j^i + 1$ and $\beta_1 = (R - x_j^i) + \beta_0 = (R - x_j^i) + 1$. As can be seen in Equation 3.16, the posterior distribution is indeed only dependent on the data.

3.3.3 Bayes factor

We will use the Bayes factor to evaluate informative hypotheses. A Bayes factor (BF) is commonly represented as the ratio of the marginal likelihoods of two hypotheses (Kass & Raftery, 1995). Klugkist et al. (2005) and Hoijtink (2012, 2012, p. 51–52, 57–59) show that for inequality constrained hypotheses of the form presented in Equation 3.3 the ratio of marginal likelihoods expressing support for H_m^i relative to H_u^i can be rewritten as

$$BF_{mu}^i = \frac{f_m^i}{c_m^i}. \tag{3.17}$$

The Bayes factor balances the relative fit and complexity of two hypotheses. Fit and complexity are called relative because they are relative with respect to the unconstrained hypothesis. In the remainder of this text, referrals to fit and complexity should be read as *relative* fit and complexity. The complexity c_m^i is the proportion of the unconstrained prior distribution for H_u^i in agreement with H_m^i

$$c_m^i = \int_{\boldsymbol{\pi}^i \in H_m^i} h(\boldsymbol{\pi}^i \mid H_u^i) \delta \boldsymbol{\pi}^i. \tag{3.18}$$

Using Equation 3.15 with $\alpha_0 = \beta_0 = 1$ for each $\boldsymbol{\pi}^i$ it is ensured that the prior distribution is unbiased with respect to hypotheses that belong to an equivalent set. Consider for example, $H_1 : \pi_1 > \pi_2 > \pi_3 > \pi_4$ and $H_2 : \pi_1 > \pi_2 > \pi_4 > \pi_3$. These hypotheses, and the other 22 possible ordering of $\boldsymbol{\pi}^i$, are equally complex and should thus have the same complexity. Using Equation 3.15, this complexity is computed as $\frac{1}{24}$ for each of the set of 24 equivalent hypotheses (Hoijtink, 2012, p. 60).

The fit f_m^i is the proportion of the unconstrained posterior distribution in agreement with H_m^i:

$$f_m^i = \int_{\boldsymbol{\pi}^i \in H_m^i} g(\boldsymbol{\pi}^i \mid \mathbf{x}^i, H_u^i) \delta \boldsymbol{\pi}^i. \tag{3.19}$$

The appendix describes how stable estimates of the complexity and fit can be computed using MCMC samples from the prior and posterior distribution, respectively.

Since Equation 3.17 is a ratio of two marginal likelihoods (one for H_m^i and one for H_u^i) it follows that

$$BF_{mm'}^i = \frac{BF_{mu}^i}{BF_{m'u}^i} = \frac{f_m^i / c_m^i}{f_{m'}^i / c_{m'}^i}, \tag{3.20}$$

and that

$$BF_{m\not m}^i = \frac{f_m^i / c_m^i}{f_{\not m}^i / c_{\not m}^i} = \frac{f_m^i / c_m^i}{1 - f_m^i / 1 - c_m^i}. \tag{3.21}$$

Three hypothetical $N = 1$ datasets with $J = 4$ and $R = 7$ are presented in Table 3.1. Three possible informative hypotheses regarding these data are H_1^i from Equation 3.5, $H_{\not 1}^i$ and H_2^i from Equation 3.7. The table presents the complexity, fit and Bayes factors of these hypotheses. As can be seen in the table, the complexity of H_1^i is $.04 = 1/24$ and $c_2^i = .5$. The table illustrates that complexity depends on the hypotheses but not on the data: for each of the three data examples the complexities are the same.

The first example (Person 1) in Table 3.1 contains data that are in agreement with H_1^i , and therefore also with H_2^i, since H_1^i is a specific case of H_2^i. This is reflected by $f_1^1 = .556$ and $f_2^1 = .996$. Because H_1^i is quite specific, it can easily conflict with the data. For example, based on $x_2^1 = 5$ and $x_3^1 = 4$, it is not very certain that $\pi_2^1 > \pi_3^1$. In contrast, H_2^i is less specific, does not involve the constraint $\pi_2^1 > \pi_3^1$, and therefore f_2^1 is larger than f_1^1. Bayes factors balance complexity and fit of the hypotheses, resulting in $BF_{1u}^1 = 13.16$, $BF_{2u}^1 = 2.00$, $BF_{12}^1 = 6.59$ and $BF_{2\not 2}^1 = 99$. Interpreting the size of Bayes factors is a matter that needs some discussion. Firstly, it is important to distinguish the different interpretations of BF_{mu}^i, $BF_{mm'}^i$ and $BF_{m\not m}^i$. In itself, BF_{mu}^i represents the relative change in the support

Table 3.1
Complexity, fit, and Bayes factors for three hypothetical N = 1 studies.

i	x_1^i	x_2^i	x_3^i	x_4^i	c_1^i	c_2^i	f_1^i	f_2^i	BF_{1u}^i	BF_{2u}^i	$BF^i_{1/1}$	BF_{12}^i	$BF^i_{2/1}$
1	7	5	4	1	.04	.50	.56	.99	13.16	2.00	28.39	6.59	99
2	7	2	5	1	.04	.50	.06	.89	1.40	1.79	1.43	.78	8.09
3	3	4	6	1	.04	.50	.01	.51	.24	1.01	.23	.24	1.04

Note. $H_1^i = \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i$ and $H_2^i = \frac{\pi_1^i + \pi_2^i}{2} > \frac{\pi_3^i + \pi_4^i}{2}$.

for H_m^i and H_u^i caused by the data. For example, in Table 3.1 we find that the belief for H_1^1 has increased 13 times and the belief for H_2^1 has increased 2 times. This shows that, although with varying degrees, both hypotheses are supported by the data. If we compute $BF_{mm'}^i$ we can quantify the relative change in support for H_m^i and $H_{m'}^i$ caused by the data. For example, $BF_{12}^1 = 6.6$, indicating that the relative support for H_1^1 compared to H_2^1 has increased by a factor 6.6. However, BF_{12}^1 is only a relative measure of support, that is, the best of the hypotheses involved may still be an inadequate representation of the within person population that generated the data. Note that BF_{mu}^1 and $BF^i_{m/m}$ are always both larger or smaller than 1. However, by definition BF_{mu}^i ranges from 0 to $c_m^{i\,-1}\,\frac{1}{c_m^i}$ and $BF^i_{m/m}$ ranges from 0 to infinity. Therefore, we prefer to interpret the latter to determine if the best of a set of hypotheses is also a good hypothesis. By computing $BF^i_{m/m}$, we can determine whether the best hypothesis, in this case H_m^i, is also a good hypothesis, because we get an answer to the question 'is or isn't H_m^i supported by the data?'. In Table 3.1, $BF^i_{1/1} = 28.4$ indicates that the data caused an increase in believe for H_m^i compared to $H^i_{m/}$, which implies that it is a good hypothesis. Note that this does not rule out the possibility of other, perhaps better, good hypotheses.

A second issue is the interpretation of the strength of Bayes factors. Although some guidelines have been provided (interpret 3 as the demarcation for the size of BF_{ab}, providing marginal and positive evidence in favor of H_a, e.g. Kass & Raftery, 1995), we choose not to follow them. In the spirit of a famous quote from (Rosnow & Rosenthal, 1989), 'surely God loves a BF of 2.9 just as much as a BF of 3.1', we want to stay away from cut-off values in order not to provide unnecessary incentives for publication bias and sloppy science (Konijn, Van de Schoot, Winter, & Ferguson, 2015). In our opinion, claiming that a Bayes factor of 1.5 is not very strong evidence and that a Bayes factor of 100 is strong evidence will not result in much debate. It is somewhere between those values that scientists may disagree about the strength. In this paper we used the following strategy to decide when a hypothesis can be considered best for a person: a hypothesis m is considered the best of a set of M hypotheses if the evidence for H_m is at least $M - 1$ times (with a minimum value of 2) stronger than for any other hypothesis m'. This requirement ensures that the posterior probability for the best hypothesis is at least .5 if all hypotheses are equally likely a priori. For example, if two hypotheses are considered, one should be at least 2 times more preferred than the other, resulting in posterior probabilities of at least .66 versus .33. If three hypotheses are considered, the resulting posterior probabilities will be at least .50 versus .25 and .25, which corresponds to a twofold preference of one hypothesis over both alternatives. For four hypotheses the posterior probabilities should be at least .50 versus

.16, .16 and .16, corresponding to relative support of at least 3 times more for the best hypothesis than for any other hypothesis. Note that, although these choices seem reasonable to us, other strategies can be thought of and justified.

For Person 2 in Table 3.1 H_2^i has gained slightly more belief than H_1^i, since $BF_{12}^2 = .78$ ($BF_{21}^2 = 1.28$). Based on this Bayes factor, H_2^i is not convincingly the better hypothesis of the two. It is important to note that Bayes factors for different persons do not necessarily express support in favor of one or the other hypothesis. It is very possible that Bayes factors for different persons are indecisive. Looking at $BF_{1\not1}^2 = 1.43$ and $BF_{2\not2}^2 = 8.09$, H_2^i seems quite a good hypothesis, whereas H_1^i is not much more supported than its complement. Finally, Person 3 in Table 3.1 shows data that do not seem to be in line with either H_1^i or H_2^i. According to $BF_{1u}^3 = .24$, the support for H_1^3 relative to H_u^3 has decreased after observing the data. According to $BF_{2u}^3 = 1.01$, the data do not cause a change in support for H_2^3 relative to the unconstrained hypothesis. When we look at $BF_{12}^3 = .24$ ($BF_{21}^3 = 4.17$), we find that H_2^i is a somewhat better hypothesis than H_1^i. However, $BF_{2\not2}^3 = 1.04$, indicating that although H_2^i is better than H_1^i, it is not a very good hypothesis. The examples in Table 3.1 show the variety in conclusions that can be obtained. There may or may not be a best hypothesis, and the best hypothesis may or may not be a good hypothesis.

3.3.4 Illustration

For Zedelius et al. (2011), the main goal was to select the best hypothesis from H_1^i, H_2^i, H_3^i and H_4^i presented in Equations 3.10, 3.11, 3.12 and 3.13. The Bayes factors presented in the first four columns of Table 3.2 can be used to select the best hypothesis for each person. If a best hypothesis is selected, it is also of interest to determine whether this hypothesis is a good hypothesis. The last four columns of Table 3.2 can be used to determine whether the best hypothesis is also 'good'.

For Person 1, H_3^1 is $1.98/.59 \approx 3.36$ times more supported than H_1^1, $1.98/.93 \approx 2.13$ times more supported than H_2^1 and $1.98/.26 \approx 7.62$ times more supported than H_4^1. Although H_3^1 is more supported than the other three hypotheses, a Bayes factor of 2.13 does not seem very convincing. Comparing the relative strength of the support for all informative hypotheses for Person 1 leaves us with the conclusion that no single best hypothesis could be detected. This implies that for Person 1, we would not be quite certain which hypothesis best describes the data Thus, we may conclude that for Person 1, it is difficult to select a best hypothesis.

For Person 8, none of the informative hypotheses is preferred over the unconstrained hypothesis. Thus, for each of the formulated hypotheses, our belief has decreased after obtaining the data. If we have to select a best hypothesis, however H_2^8 and H_4^8 are respectively $.16/.03 \approx 5.3$ and $.19/.03 \approx 6.3$ times more supported than H_3^8 and at least $.16/.01 \approx .19/.01 \approx 17$ times more supported than H_1^8. However, based on $BF_{2\not2}^8 = .15$ and $BF_{4\not4}^8 = .10$ we can conclude that although H_2^8 and H_4^8 are convincingly preferred over the other two hypotheses, neither is a good hypothesis for this person.

For Person 14, H_2^{14} is $6.53/.55 \approx 11.9$ times more supported than H_1^{14}, $6.53/.56 \approx 11.7$ times more supported than H_3^{14} and $6.53/.78 \approx 8.4$ times more supported than H_4^{14}. We find that $BF_{2\not2}^{14} = 8.61$, so besides the fact that H_3^{14} is preferred over the other hypotheses it

is a good hypothesis, too. Thus, we may conclude that for Person 14 we can find a best hypothesis that appears to be a good hypothesis, too.

For Person 20, H_4^{20} is at least 79 times more supported than H_1^{20}, H_2^{20} and H_3^{20}. Thus, H_4^{20} is the best hypothesis from the set. However, because $BF_{4\not}^{20} = .65$ we can conclude that even though H_4^{20} was the best hypothesis, it is not a good description of the data.

These examples show that it differs per person whether a best hypothesis can be detected, which hypothesis this is, and how strong the evidence is relative to the other hypotheses. Based on Table 3.2, Zedelius et al. (2011) can conclude for each individual what the best hypothesis is, and whether it is a good hypothesis. We find that the sample contains persons for whom a best hypothesis can be detected, but this hypothesis is not a good hypothesis (Persons 20 and 21). Additionally, there are individuals for whom a best hypothesis can be detected and the best hypothesis is good (Persons 6, 14, 15, 16, 17, 19, 22 and 23). For the remaining individuals, no best hypothesis could be selected. Someone else evaluating these Bayes factors might come to slightly different conclusions, if they apply a different rule to decide what makes a hypothesis the best from a set.

The second goal of this paper was to determine whether the sample of individuals comes from a homogeneous population with respect to the support for the hypotheses of interest. The first impression gained from Table 3.2 is that this is not the case. However, this topic will be pursued in depth in the next section.

3.4 A P-population of WP-populations

Looking at the Bayes factors in Table 3.2 is a rather ad hoc manner to answer the question whether the sample comes from a population that is homogeneous in its support for the the hypotheses under consideration and which hypothesis is the best. By aggregating the individual Bayes factors we can try to evaluate in more detail to what extent individuals are homogeneous with respect to a hypothesis. If H_m^i is evaluated for P independent persons the corresponding individual Bayes factors can be multiplied into a P-population Bayes factor (Stephan & Penny, 2007):

$$\text{P-BF}_{mu} = \prod_{i=1}^{P} \text{BF}_{mu}^{i}, \tag{3.22}$$

which expresses the support for H_m relative to H_u, where

$$H_m = H_m^1 \cup ... \cup H_m^P, \tag{3.23}$$

which states that H_m^i holds for every person $i = 1, ..., P$, and

$$H_u = H_u^1 \cup ... \cup H_u^P, \tag{3.24}$$

which is the union of H_u^i for $i = 1, ..., P$. In this section using the Bayes factor, H_m^i and H_m are compared with H_u^i and H_u, respectively. However, analogously, H_u^i could be replaced by $H_{m'}^i$ or $H_{\not m}^i$ rendering P-$BF_{mm'}$ and P-$BF_{m\not m}$, respectively. Note, that this is *not* the Bayes factor describing the relative evidence for Hm and Hm' with regard to the P-population

Table 3.2
Individual Bayes factors for the Zedelius (2011) data

i	BF^i_{1u}	BF^i_{2u}	BF^i_{3u}	BF^i_{4u}	$BF^i_{1/u}$	$BF^i_{2/u}$	$BF^i_{3/u}$	$BF^i_{4/u}$
1	0.59	0.93	1.98	0.26	0.59	0.93	2.06	0.15
2	3.33	1.49	4.67	0.45	3.33	1.52	5.54	0.29
3	1.02	1.31	1.63	1.41	1.02	1.33	1.68	2.37
4	0.03	0.10	0.58	1.22	0.03	0.10	0.57	1.55
5	3.79	2.39	4.92	1.02	3.79	2.55	5.91	1.04
6	543.90	17.95	13.74	1.43	551.21	68.72	30.30	2.51
7	1.44	3.45	2.88	1.23	1.44	3.87	3.14	1.58
8	<0.01	0.16	0.02	0.19	<0.01	0.15	0.02	0.10
9	3.06	6.16	3.25	1.94	3.06	7.95	3.59	30.74
10	2.60	3.41	2.75	0.99	2.60	3.81	2.97	0.97
11	0.05	0.24	0.55	1.21	0.05	0.23	0.54	1.53
12	1.29	1.70	1.55	0.44	1.29	1.76	1.58	0.28
13	0.30	3.50	2.66	0.79	0.30	3.93	2.86	0.65
14	0.55	6.53	0.56	0.78	0.55	8.61	0.55	0.64
15	21.84	2.01	6.41	1.73	21.85	2.10	8.35	6.28
16	0.18	0.45	3.21	1.22	0.18	0.44	3.54	1.56
17	22.30	5.15	3.88	1.91	22.31	6.28	4.42	20.64
18	0.32	1.37	0.55	0.62	0.32	1.39	0.54	0.45
19	<0.01	<0.01	0.03	1.96	<0.01	<0.01	0.03	40.41
20	<0.01	<0.01	0.01	0.79	<0.01	<0.01	0.01	0.65
21	0.09	0.41	0.40	1.43	0.09	0.40	0.39	2.50
22	15.78	5.59	4.82	1.58	15.78	6.98	5.77	3.68
23	20.92	4.39	7.62	1.60	20.93	5.15	10.64	3.92
24	0.15	1.16	0.32	1.01	0.15	1.17	0.31	1.02
25	7.21	3.16	3.26	0.76	7.21	3.49	3.61	0.61
26	0.06	0.13	0.38	0.58	0.06	0.13	0.37	0.41

Note. H^i_1 , H^i_2 and H^i_3 (Equations 3.10–3.12) are evaluated against H^i_u and their complement.

parameters $\boldsymbol{\pi}$. Individual data *could* be used to evaluate a Bayes factor with respect to the P-population $\boldsymbol{\pi}$, but our focus here is on the collection of individual WP-populations $\boldsymbol{\pi}^i$. Another way to interpret this P-BF is in the context of *synthesis* of knowledge with respect to the individual evaluated hypotheses H^i_m. Thus, it is a measure of the extent to which a hypothesis holds for every individual, rather than on average.

Table 3.3 shows seven hypothetical sets of six individual Bayes factors comparing H^i_m to H^i_u . The P-BF is presented for each set. For example, Set 1 results in a P-BF of 64, indicating that it is 64 times more likely that H^i_m holds for all persons i, than that it does not hold for all persons. However, the table shows an undesirable property of P-BF, namely that it is a function of P. As can be seen, both in Set 1, 2 and 3, the P-BF is 64. Nevertheless, it is clear that all individual Bayes factors in Set 1 express stronger evidence than in Sets 2 and 3.

Stephan & Penny (2007) have suggested using the geometric mean of the product of individual Bayes factors to render a summary that is independent of P:

$$\text{gP-BF}_{mu} = \sqrt[P]{\text{P-BF}_{mu}}, \tag{3.25}$$

which is a measure of the 'average' support in favor of H_m relative to H_u found in P persons.

In other words, it can be interpreted as the Bayes factor that is expected for the $P + 1^{st}$ individual sampled from the P-population.

As can be seen in Table 3.3, the gP-BF$_{mu}$ does not depend on P. For example, in Set 1 the gP-BF is 8.00 and in the larger Sets 2 and 3, the average support for H_m is 2.83 and 2.00, respectively, while the P-$BF_{mu} = 64$ for each of these sets.

If multiple hypotheses are considered, gP-$BF_{mm'}$ and gP-$BF_{m/m}$ can be derived similar as $BF^i_{mm'}$ and $BF^i_{m/m}$. It is important to keep in mind that the gP-BF_{mu} is a summary measure and does not have the same properties as individual Bayes factors. Such a property is that BF^i_{mu} and $BF^i_{m/m}$ are always both smaller or larger than 1. For example, if $BF^1_{1u} = 0.2$, then $BF^1_{1/1} = 0.4$, and if $BF^2_{1u} = 1.8$ then $BF^2_{1/1} = 9$. This is not true for gP-BF_{mu} and gP-$BF_{m/m}$. To continue the example based on the Bayes factors for persons 1 and 2, gP-$BF_{1u} = 0.6$ and gP-$BF_{1/1} = 2$. For interpretation of the gP-BF, it is important to keep in mind that gP-BF_{mu} is a summary of all BF^i_{mu}, and thus cannot be translated into gP-$BF_{m/m}$, which is a summary of all $BF^i_{m/m}$. Note that if a switch in direction occurs, both geometric Bayes factors are generally both close to 1, therefore not causing any very contradicting conclusions.

However, the gP-BF$_{mu}$ has another issue. Table 3.3 shows that different sets of individual Bayes factors can lead to the same gP-BF$_{mu}$. For example, in Sets 3, 4 and 5 the same gP-BF is obtained. Set 3 contains only Bayes factors that are close to the gP-BF = 2 and all support H^i_m. Set 4 seems similar in the strength of support in the individual Bayes factors, although there seems to be more variation than in Set 3, and we find one Bayes factor that does not support H^i_m. Finally, Set 5 contains four Bayes factors that express support for H^i_u over H^i_m, while two Bayes factors express relatively strong support in favor of H^i_m over H^i_u. The fact that the Bayes factors from Sets 3 and 4 come from populations that are more homogeneous in their preference for H^i_u than Set 5 is not represented well by the gP-BF$_{mu}$. Therefore, an additional measure, the evidence rate (ER$_{mu}$), is introduced that describes the consistency in the preferred hypothesis in multiple individual Bayes factors:

$$ER_{mu} = \begin{array}{ll} \frac{1}{P}\sum_{i=1}^{P} I_{BF^i_{mu}<1} & \text{if} \quad \text{gP-BF}_{mu} < 1 \\ \frac{1}{P}\sum_{i=1}^{P} I_{BF^i_{mu}>1} & \text{if} \quad \text{gP-BF}_{mu} > 1 \end{array}, \tag{3.26}$$

where $I_{BF^i_{mu}>1} = 1$ if $BF^i_{mu} > 1$ and 0 otherwise. Thus, the ER_{mu} is the proportion of individual BF^i_{mu} that expresses support for H^i_m or for H^i_u if the gP-BF$_{mu}$ expresses support for H_m or H_u, respectively. For example, if gP-BF$_{mu} > 1$, an ER_{mu} of 1 indicates that all individual Bayes factors express support for H^i_m. An ER of .5, indicates that 50% of the individual Bayes factors expresses support for H^i_m, and 50% expresses support for H^i_u. An ER close to 1 indicates homogeneity among the individual Bayes factors. The lower the ER, the stronger the evidence that the ordering of the individual success probabilities are not homogeneous with respect to the hypotheses under consideration. Looking at Table 3.3, we find that in Set 3 all individual Bayes factors support H^i_m, this is reflected in an $ER_{mu} = 1$. In Set 4 most, but not all individual Bayes factors support H^i_m, resulting in $ER_{mu} = .83$. This implies that there is no perfect homogeneity among the individual Bayes factors. Finally, in Set 5, four of six individual Bayes factors support H^i_u, while gP-BF$_{mu}$ supports H^i_m. The ER_{mu} of .33 indicates that Set 5 is not likely to come from a homogeneous population with respect to the hypotheses under consideration.

There is still one issue that needs to be resolved. Set 6 and 7 result in the same gP-BF$_{mu}$

and ER_{mu} as Set 3, but are not similar in individual contributions. Set 6 contains an outlier that expresses strong evidence for H^i_m, whereas all other cases express only weak support for H^i_m. Without this outlier, the gP-BF_{mu} would be much lower. Set 7 contains two Bayes factors that express very little support for H^i_m, whereas the other four cases express stronger support for H^i_m. Without these two 'weak' cases, the gP-BF_{mu} would be somewhat higher. In contrast, Set 3 contains Bayes factors that are rather constant around gP-BF, removing any of these cases would not affect the gP-BF_{mu} too much. To describe presence and direction of skewness among individual Bayes factors with respect to the gP-BF_{mu}, a final measure is introduced: the stability rate.

The stability rate (SR_{mu}) is a measure of skewness among individual Bayes factors with respect to the gP-BF_{mu}. It can be written as:

$$SR_{mu} = \begin{array}{ll} \frac{1}{P}\sum_{i=1}^{P} I_{BF^i_{mu}<\text{gP-BF}_{mu}} & \text{if} \quad \text{gP-BF}_{mu} < 1 \\ \frac{1}{P}\sum_{i=1}^{P} I_{BF^i_{mu}>\text{gP-BF}_{mu}} & \text{if} \quad \text{gP-BF}_{mu} > 1 \end{array}, \tag{3.27}$$

where $I_{BF^i_{mu}<\text{gP-BF}_{mu}} = 1$ if $BF^i_{mu} < \text{gP-BF}_{mu}$ and 0 otherwise. The SR_{mu} describes the proportion of individual Bayes factors that expresses support stronger than the gP-BF for the hypothesis preferred by gP-BF_{mu}. In Sets 1, 2, 3 and 4 of Table 3.3 the gP-BF_{mu} prefers H^i_m over H^i_u. Individual Bayes factors that express stronger support for H^i_m than gP-BF are presented in bold in the table. For each of these sets, the $\text{SR}_{mu} = .50$, indicating that half of the individual Bayes factors expresses support for H^i_m stronger than gP-BF. The other half expresses support either for H^i_u or weaker support for H^i_m. An SR_{mu} close to .50 indicates that the individual Bayes factors are evenly distributed around gP-BF.

An SR_{mu} smaller than .50, as in Set 5 and 6, indicates that less than half of the individual Bayes factors express stronger support for H^i_m than gP-BF. Consequently, the gP-BF_{mu} is relatively large because of a minority of individual Bayes factors that are relatively large. The gP-BF_{mu} is overestimated because of this minority. In Set 5 the gP-BF_{mu} supports H^i_m, while the majority of individual Bayes factors support H^i_u. The gP-BF_{mu} is no longer a representative 'average' support. Reversely, an SR_{mu} larger than .50 indicates that only relatively few individual Bayes factors express weaker support than gP-BF (see Set 7). Thus, for $\text{SR}_{mu} > .50$, the gP-BF_{mu} is relatively close to 1 because of a minority of individual Bayes factors that express support that is relatively weak. As an effect, the strength of support is underestimated.

Thus, the gP-BF_{mu} can be used to express the average support of the individual Bayes factors. In order to assess whether the individual Bayes factors come from a homogeneous population, the ER_{mu} can be used. A high evidence rate indicates high agreement in preferred hypothesis among individual Bayes factors, and thus more homogeneity. Finally, the SR_{mu} gives an indication of how the individual Bayes factors are distributed around the gP-BF_{mu}. Note that the equations presented for the ER and SR describe those corresponding to gP-BF_{mu}. If the interest is in gP-$\text{BF}_{mm'}$ or gP-$\text{BF}_{m'm}$, the ER and SR should be computed using the individual $BF^i_{mm'}$s and $BF^i_{m'm}$s. The individual Bayes factors are the relevant quantities in the ER and SR, and therefore these should be used.

Table 3.3
Hypothetical individual Bayes factors (P = 6), gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
BF^1_{mu}	**9.00**	**3.20**	1.40	0.80	0.90	**6.40**	1.01
BF^2_{mu}	7.11	2.70	**2.70**	1.50	0.93	1.40	1.30
BF^3_{mu}	-	2.30	1.80	**2.50**	0.85	1.80	**2.50**
BF^4_{mu}	-	**3.22**	**2.10**	**4.33**	0.88	1.40	**3.10**
BF^5_{mu}	-	-	1.60	**3.10**	**6.30**	1.60	**2.60**
BF^6_{mu}	-	-	**2.80**	1.59	**16.23**	1.77	**2.42**
P-BF_{mu}	64.00	64.00	64.00	64.00	64.00	64.00	64.00
gP-BF_{mu}	8.00	2.83	2.00	2.00	2.00	2.00	2.00
ER_{mu}	1	1	1	.83	.33	1	1
SR_{mu}	.50	.50	.50	.50	.33	.17	.67

3.4.1 Illustration

Using the individual Bayes factors presented in Table 3.2 the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ can be computed for the data of Zedelius et al. (2011). The first row of Table 3.4 gives the gP-BF$_{mu}$ based on the individual Bayes factors from Table 3.2. The ER$_{mu}$ and SR$_{mu}$ are presented in the second and third row. Based on the gP-BF$_{mu}$ we can conclude that H_3 receives approximately $1.125/.510 \approx 2.21$ times more support than H_1, and only about $1.125/.910 \approx 1.125/.949 \approx 1.2$ times more support than H_2 and H_4. Thus, H_3 is somewhat preferred over H_1, but cannot be distinguished from H_2 and H_4. Furthermore, since gP-BF$_{2u}$ = 1.014, gP-BF$_{3u}$ = 1.235 and gP-BF$_{4u}$ = 1.412, it can be concluded that none of the hypotheses is convincingly the best description for all individuals and none of the hypotheses are clearly a better description of all individuals than their complement is.

Additionally, we find that the ER$_{mu}$ for the comparison of H_1 with H_u is .500, indicating that approximately half of the individual Bayes factors expresses support for H^i_1, while the other half expresses support for H^i_u. Similarly, ER$_{2u}$, ER$_{3u}$ and ER$_{4u}$ are .346, .615 and .423 indicating that for these hypotheses, too, there is little homogeneity among the individual Bayes factors. Only SR$_{1u}$ is rather close to .50, and consequently, it is not likely that the gP-BF$_{mu}$ is affected by one or more influential cases having a (much) smaller BF than the majority. For the other hypotheses, there is indication that the strength of the gP-BF$_{mu}$ is affected by skewness among the individual Bayes factors.

Based on the gP-BF$_{mu}$, ER$_{mu}$, and SR$_{mu}$, we can draw the following conclusions. Firstly, using the gP-BF$_{mu}$ no hypothesis could be selected as the best hypothesis from the set. The SR$_{mu}$s indicate that for all hypotheses but H^i_1 imbalance among individual Bayes factors was present. Furthermore, the relatively low ER$_{mu}$s indicate that it is unlikely that the individuals come from a homogeneous population with respect to any of the specified hypotheses. Finally, none of the hypotheses appears to be a good description of the ordering of the individual success probabilities. Thus, based on these findings it seems unlikely the P-population is homogeneous with respect to the WP-population hypotheses that were considered.

A within-person experiment, such as conducted by Zedelius et al. (2011), is quite common in social and neuro-psychological research. The theory and hypotheses for these experiments are often at the WP-population level. Examples are Moreland & Zajonc (1982), who wonder *"... whether mere exposure to other people [...] is a sufficient condition for the enhancement of their perceived similarity to ourselves."* (p. 397) and Klimecki et al. (2016), who hypothesize that *"... altruistic motivation is elicited by empathy felt for a person in need."* (p. 1). Zedelius et al. (2011) write that *"... rewards cause people to invest more effort in a task..."*, *"... the intriguing hypothesis that [...] reflective thoughts hinder ongoing performance..."* (p. 355) and *"... participants performed significantly better..."* (p.356). These fragments contain theory or expectations regarding the behavior of individual people.

Although WP-population hypotheses are formulated, the analyses are usually executed at the P-population level. In the original Zedelius et al. (2011) paper, the data were analyzed by means of a repeated measures ANOVA, which tests differences in the P-population means. The conclusions obtained from this analysis imply that H_2 holds at the P-population level. Often the, usually implicit, assumption is that if a hypothesis holds at the P-population level, it holds for all individuals. The current analysis shows that although H_2 is a reasonable hypothesis at the P-population level, it appears not to be the single best hypothesis under consideration and is not a good hypothesis for all individuals. The assumption that an average conclusion holds for all individuals is in this case violated. It is important that psychological researchers are aware of the fact that conclusions at the P-population level cannot be transferred to the individual level without testing this. Within-person experiments offer rich data that allow for the evaluation of individual hypotheses, through which the assumption that a hypothesis holds for everyone can be tested. This paper introduces an approach with which this can be done.

Table 3.4
The gP-BF, ER and SR for the data of Zedelius et al. (2011).

	BF_{1u}	BF_{2u}	BF_{3u}	BF_{4u}	$BF_{1/1}$	$BF_{2/2}$	$BF_{3/3}$	$BF_{4/4}$
gP-BF	0.510	0.910	1.125	0.949	0.511	1.014	1.235	1.412
ER	0.500	0.346	0.615	0.423	0.500	0.654	0.615	0.577
SR	0.423	0.308	0.615	0.385	0.423	0.654	0.615	0.500

*Note.*The hypotheses evaluated are H_1^i, H_2^i, H_3^i and H_4^i as in Equations 3.10–3.13 and their complement.

3.5 Determining the sample size and number of replications for a study

Say, a researcher has a research question that he wants to test by means of an experiment. This research question defines which and how many conditions J should be considered and results in one or multiple hypotheses of interest. The researcher is then left with two choices regarding the experiment, namely, the number of replications R used in each trial and the sample size P. This section will describe a method to choose R and P.

In the previous section, a method to evaluate a set of individual Bayes factors has been introduced in the form of three measures: gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$. It is important to investigate the properties of these measures as a function of sample size and the number of replications. In other words, if indeed all individuals are homogeneous with respect to an individual informative hypothesis, which are the sample size and number of replications required for gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ to succeed in detecting this and, analogously, if individuals are not homogeneous, can this be derived from these measures?

Through a sensitivity analysis it can be determined for which sample size and number of replications the gP-BF$_{mu}$ can be expected to prefer the hypothesis that is in agreement with the true P-population, the ER$_{mu}$ is sufficiently high and SR$_{mu}$ is close to .5. The choice for what values the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ behave as desired is subjective. In line with our reasoning for the interpretation of individual Bayes factors as described on page 55, the choice for when the strength of support in gP-BF is sufficient to prefer one hypothesis over another is subjective and no guidelines are provided. Additionally, we will consider .9 to be sufficiently high for the ER$_{mu}$, that is, a maximum 10% of individual Bayes factors prefers a different hypothesis than the majority, and a .1 margin around .5 to be reasonable for the SR$_{mu}$, that is, the proportion of individual Bayes factors expressing stronger support than gP-BF$_{mu}$ is between .4 and .6.

Using R (R Core Team, 2013), software has been developed with which such a study design analysis can be executed[2]. While discussing the options of this program, we focus on the evaluation of gP-$BF_{m\not m}$, in order to arrive at an appropriate study design to determine whether H^i_m holds for everyone in the P-population. The program can analogously be used for Study design analyses for gP-$BF_{mm'}$ or gP-BF_{mu}. The required input and the algorithm used are illustrated using Zedelius et al. (2011), as it could have been conducted before starting the data collection.

The R program requires as input the number of conditions J and hypotheses that a researcher wants to investigate. Additionally, the numbers of replications R and the sample sizes P that a researcher is willing to consider should be specified. Using this input, the following steps are executed:

- For each hypothesis of interest H^i_m, three P-populations are specified, one where H^i_m is true for all WP-populations, one where $H^i_{\not m}$ is true for all WP-populations and a mixture of these two populations. In the next section these P-populations are specified in more detail for the example from Zedelius et al. (2011).
- For each P-population, the program generates 10,000 WP-populations, that is, parameter vectors $\boldsymbol{\pi}^i$ of size J.
- For each R specified by the user, $\mathbf{x}^i$ is sampled from $\boldsymbol{\pi}^i$.
- For each $\mathbf{x}^i$, $BF^i_{m\not m}$ is computed.

This results in 10,000 individual Bayes factors for each combination P-population and R. For computational reasons, this set will be used as a surrogate for the true infinite P-population. For each surrogate P-population then the following steps are followed:

[2]The software with accompanying manual can be downloaded on https://github.com/fayetteklaassen/OneForAll, or be obtained by contacting the first author at klaassen.fayette@gmail.com. For assistence with or questions about the software, please also contact the first author.

- For each sample size P and number of replications R, 1000 sets of individual Bayes factors are sampled with replacement from the surrogate P-population.
- For each set, the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ are computed, resulting in 1000 values of each measure for every sample size P and number of replications R.
- From these 1000 values of gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ the 2.5, 50 and 97.5 percentiles are obtained. The 50 percentile, the median, is used to summarize what values can be expected for each of these measures. The desired values of these expectations are, as described above subjectively defined, for the gP-BF$_{mu}$, above .9 for the ER$_{mu}$ and within a .1 margin from .5 for the SR. The 2.5 and 97.5 percentiles indicate the range in which 95% of the sampled gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ lay. If this range is very wide and includes non-reasonable values the combination of R and P might not be appropriate even when the expected value is of a desired level. In the next section we will illustrate how this information can be used to determine the R and P required to execute a study.

3.5.1 Illustration

This section describes a sensitivity analysis for the determination of the number of replications R and sample size P, where the setup of Zedelius et al. (2011) will be used as starting point. Of course, such an analysis should be executed prior to the data collection, which was already done by Zedelius et al. (2011). However, for the illustration we will do the analysis as if no data has been collected yet. This will provide us with the knowledge whether the eventually chosen R and P were sufficient according to the sensitivity analysis. The first step of the sensitivity analysis described in the previous section requires a research question leading to the number of conditions J and a set of hypotheses representing the researchers' expectations. The research question of Zedelius et al. rendered three hypotheses, Equations 3.10–3.12, about the ordering of success probabilities in the $J = 8$ experimental conditions. For this illustration, only H_1^i as in Equation 3.2 is considered. This results in the following parameters for the sensitivity analysis:

- *Number of conditions.* Zedelius et al. (2011) considered 8 different conditions, so $J = 8$.
- *Hypothesis.* The hypothesis that will be considered for this illustration is H_1^i . From this hypothesis, three relevant P-populations are derived.

  P-population 1. In this P-population all individuals adhere to H_1^i . Using this population the median values of the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ can be determined if H_m^i holds for everyone. To compute these median values the individual parameters $\boldsymbol{\pi}^i$ are repeatedly sampled from the prior distribution under H_1^i:

$$h(\boldsymbol{\pi}^i|H_1^i) \propto h(\boldsymbol{\pi}^i|H_u^i) I_{\boldsymbol{\pi}^i \in H_1^i}, \tag{3.28}$$

  where $I_{\boldsymbol{\pi}^i \in H_1^i} = 1$ if $\boldsymbol{\pi}^i$ is in agreement with H_1^i and 0 otherwise.

  P-population 2. In this P-population all individuals adhere to H_1^i. Using this population the expected values of the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ can be determined

if $H^i_{\not m}$ holds for everyone. The individual parameters $\boldsymbol{\pi}^i$ are sampled from the prior distribution under $H^i_{\not 1}$, that is:

$$h(\boldsymbol{\pi}^i|H^i_{\not 1}) \propto h(\boldsymbol{\pi}^i|H^i_u) I_{\boldsymbol{\pi}^i \in H^i_{\not 1}}, \tag{3.29}$$

where $I_{\boldsymbol{\pi}^i \in H^i_{\not 1}} = 1$ if $\boldsymbol{\pi}^i$ is in agreement with $H^i_{\not 1}$ and 0 otherwise.

P-population 3. For the third P-population, a mixture of P-population 1 and 2 is considered. Using this population the expected values of the gP-BF, ER and SR can be determined if H^i_m holds for a proportion θ of individuals in the P-populations, and $H^j_{\not m}$ holds for a proportion $1 - \theta$ of individuals. The individual parameters $\boldsymbol{\pi}^i$ are sampled from Equation 3.28 if u^i, sampled from $U(0, 1)$ is smaller than or equal to the specified proportion θ, and sampled from Equation 3.29 if u^i is larger than θ:

$$\boldsymbol{\pi}^i \sim \begin{array}{ll} h(\boldsymbol{\pi}^i|H^i_1) & \text{if} \quad u^i \leq \theta \\ h(\boldsymbol{\pi}^i|H^i_{\not 1}) & \text{if} \quad u^i > \theta \end{array}. \tag{3.30}$$

The proportion θ is set to .5, thus half of all individuals adheres to H^i_1 and the other half adheres to $H^i_{\not 1}$.

Next, the sample sizes P and number of replications R that the researchers want to consider should be chosen. Based on the choices made by Zedelius et al. (2011), the following values for P and R are considered for the sensitivity analysis:

- *Number of replications.* Zedelius et al. (2011) used 7 replications in their experiment. Additionally, it would be interesting whether more replications would result in better performance, therefore $R = 7, 14, 21$ are considered.
- *Number of individuals.* Zedelius et al. (2011) used 26 participants in their experiment. In order to mimic an a priori sensitivity analysis, the sample sizes$P = 5, 7, 10, 15, 20, 25, 30, 40, 50$ are considered.

3.5.2 Results

Figure 3.2 shows the results of the sensitivity analysis for the determination of sample size P and number of replications R. The results are presented for each of the three simulated P-populations described in the previous section. The first column of the figure shows the performance of the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ if H^i_1 is true for all individuals (P-population 1). As can be seen in the top left figure, already for small sample sizes the gP-BF$_{mu}$ expresses strong support for H_1: the lower 2.5 percentile of the gP-BF$_{mu}$ is larger then 10 for $R > 7$ and $P > 5$. The lower 2.5th percentile of the ER$_{mu}$ only stabilizes above .9 for $R = 7$ and $P > 30$ and for $R = 14, 21$, this is already achieved for $P > 10$. Stated otherwise, if H^i_1 holds for all individuals, for samples larger than 30 it is likely that less then 10 per cent of individual Bayes factors express support for $H^i_{\not 1}$. Finally, the bottom panel shows that the SR$_{mu}$ stabilizes around .55, reflecting that it is reasonable to expect slightly more than half of the individual Bayes factors to express stronger support than gP-BF$_{mu}$. This implies that the gP-BF$_{mu}$ is, on average, slightly more influenced by the 'weaker' and

contradicting individual Bayes factors. The 2.5 and 97.5 percentiles are within a margin of .1 from the median gP-BF for $P > 25$. Furthermore, we see that from around $P = 25$ the median and 2.5 and 97.5 percentiles stabilize. Thus, if H_1^i is true for all individuals, with sample size P around $25 - 30$ and $R = 7$, the gP-BF$_{mu}$ and ER$_{mu}$ perform as desired: the gP-BF$_{mu}$ shows strong evidence for the true hypothesis, the ER$_{mu}$ is high and the SR$_{mu}$ is around .5.

In the middle column of figures in Figure 3.2 $H_{\not 1}^i$ is true for all individuals. For $P > 10$ and $R > 7$, the gP-BF is smaller than .01, indicating at least 10 times more support for $H_{\not 1}^i$ than for H_1^i. As R increases, so does the median support found in the data. The lower 2.5 percentile of the ER is above .9 for $P > 30$ and $R = 14, 21$ and close to .9 for $R = 7$. The median SR is almost exactly .5 for all R for $P > 20$, and the 2.5 and 97.5 percentiles are within .1 of the median for $P > 30$. Thus, for sample sizes of 30 and larger, the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ behave as desired for $R = 7$ and even better for $R = 14, 21$.

Finally, Population 3, depicted in the right column in Figure 3.2 was chosen to be a mixture of the first two populations. Here it can be seen that if H_1^i holds for 50% of the individuals in the population, generally, $H_{\not 1}^i$ is preferred over H_1^i, although with less strength than when Population 2 was the true population. Note that this happens because it is more likely that a person coming from $h(\boldsymbol{\pi}^i|H_1^i)$ provides evidence in agreement with $H_{\not 1}^i$ than vice versa. For example, if H_1 is true but if the ordering in the data is off by one order constraint, we are likely to prefer $H_{\not 1}$. However, if one of the orderings that comprises $H_{\not 1}$ is true, a 'mistake' in one or more of the order constraints in the data does not necessarily lead to a preference for H_1, but might point to one of the other orderings under $H_{\not 1}$. The complexity of H_1 is 2.48×10^{-5} and the complexity of H_{ancel1} is $1 - 2.48 \times 10^{-5} \approx 1$. Thus, even though $\theta = .5$, $H_{\not 1}^i$ is preferred because it has a higher complexity. The ER$_{m\not 1}$ is of use here, indicating that there are multiple populations and stabilizing around .5 for $P > 30$. Although the median support found in the gP-BF$_{m\not 1}$ might indicate a preference for $H_{\not 1}^i$ over H_1^i , the ER$_{m\not 1}$ indicates inconsistency among individual Bayes factors. Finally, the median SR$_{m\not 1}$ for this population is slightly below .5, and the 2.5 and 97.5 percentiles are further than .1 from this median until P is around 40, for $R = 14, 21$ or 50 for $R = 7$. Thus, if neither of the two hypotheses hold for everyone, this is reflected in the ER$_{m\not 1}$ for every P and R that seemed reasonable if H_1^i or $H_{\not 1}^i$ were true for everyone.

Zedelius et al. (2011) eventually used 26 participants in their study and 7 replications. This is slightly lower than the suggested 30 based on the sensitivity analysis. Consulting the figures, it seems that, if H_1^i is true and $P = 26$ and $R = 7$, gP-BF$_{1\not 1}$ is expected to be between 30 and 100, the ER$_{1\not 1}$ is expected to be above .9 and the SR$_{1\not 1}$ between .43 and .67. On the other hand, if $H_{\not 1}^i$ is true for all individuals, the gP-BF$_{1\not 1}$ can be expected between 1000 and $10,000$ in support of $H_{\not 1}^i$, with the ER$_{1\not 1}$ similarly above .9 and the SR$_{1\not 1}$ between .35 and .6. Consulting the results in Table 3.4, we find that gP-BF$_{1\not 1}$= .511, ER$_{mu}$= .500 and SR$_{mu}$= .436. These results do not seem in line with either Population 1 or 2, but consulting the right column figures in Figure 3.2, they do seem in line with the mixture population. Of course, this is no evidence that indeed this mixture population with $\theta = .5$ is the most likely true P-population. However, it does indicate that even though the gP-BF$_{1\not 1}$ shows some support for $H_{\not 1}^i$ relative to H_1^i, it is not likely that $H_{\not 1}^i$ holds for everyone in the P-population.

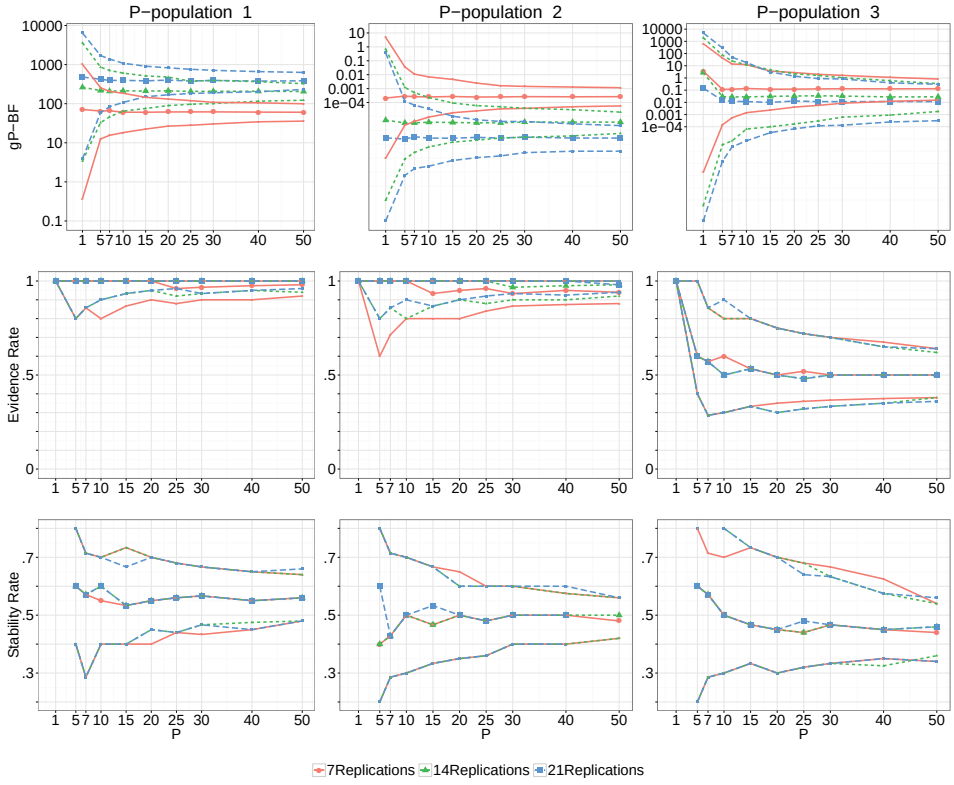


Figure 3.2. gP-BF$^i_{1l}$, ER$^i_{1l}$ and SR$^i_{1l}$ for the three generated true P-populations for $J = 8$. P-population 1 is described in Equation 3.28, P-population 2 in Equation 3.29, and P-population 3 in Equation 3.30.
Both the median and 95% interval are shown in the figures.

3.6 Discussion

After formulating within-person (WP) hypotheses, individual Bayes factors can be computed with which the support for a particular hypothesis can be derived for each person, or the best from a set of informative hypotheses can be selected. A method has been proposed to combine the individual Bayes factors of some, in order to draw conclusions for all - by answering the question whether an individual hypothesis holds for all persons in the population - and for one by determining the average support for H^i_m relative to $H^i_{m'}$ which describes what could be expected for a next individual. The geometric average of P individual Bayes factors (gP-BF) describes the average support for one hypothesis relative to another. It describes what individual Bayes factor could be expected for a next person. Together with the Evidence Rate and Stability Rate, the gP-BF can be used to assess whether one hypothesis is more supported than another for all individuals in a population. By means of a sensitivity analysis for a set of hypotheses, it can be determined for what sample size P and number of replications R in an experiment these measures behave desirable.

An R Shiny application has been developed with which a sensitivity analysis can be executed prior to data collection. By specifying hypotheses of interest, the behavior of gP-BF, ER and SR can be evaluated for various combinations of R and P. This allows researchers to collect the appropriate data for their question of interest. Besides an own sensitivity analysis, the data of the simulations used as examples in this paper can be accessed and viewed within the application. Furthermore, in the application data can be analyzed and the gP-BF, ER and SR are computed. The application and manual can be accessed on https://github.com/fayetteklaassen/OneForAll.

Chapter 4

Combining evidence over multiple individual analyses

by F. Klaassen[1]

4.1 Introduction

Hypothesis testing is omnipresent in behavioral and biomedical research, and usually concerns testing for population effects. For example, is there a difference between groups on average? This chapter presents a Bayesian method to evaluate hypotheses for each person in a sample and aggregate this result to answer the question whether a hypothesis holds for everyone in the sample, rather than on average. Using an empirical dataset, the methodology is illustrated step by step: from formulating the research question and hypotheses, to modelling the data and drawing conclusions. This chapter is structured as follows. First, informative hypotheses and Bayes factors are introduced and explained in Section 4.2. Next, a dataset and corresponding set of hypotheses is introduced in Section 4.3 that can be used for the question *Does everyone have the same best informative hypothesis?* Section 4.4 describes how individual Bayes factors can be interpreted. Section 4.5 explains how these individual Bayes factors can be combined. Throughout these sections, the methods are applied to the example dataset and hypotheses. Finally, in Section 4.6 the conclusions and limitations are discussed.

[1]In press as Klaassen, F. (in press). Combining evidence over multiple individual analyses. In R. Van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners.* Routledge.

4.2 Informative hypotheses and Bayes factors

Analysis of variance (i.e., ANOVA) and regression models are frequently used in behavioral and biomedical research. For example, consider a psychology researcher interested in the effect of interference on a memory task. The researcher plans an experiment where participants are presented a word to memorize, followed by a mask, and then asked to recall the word. The mask is a random sequence of letters (*non-word*), a word that differs by one letter from the target word (*similar word*), or a random word (*different word*). The outcome variable is reaction time. The researcher intends to test the null hypothesis $H_0 : \mu_{non-word} = \mu_{different\ word} = \mu_{similar\ word}$ that the mean reaction times in the three conditions are equal to one another against the unconstrained alternative H_a : *not* H_0, which states the expectation that at least one of the condition mean reaction times is not equal to the other conditions.

Analyzing the data by means of null hypothesis significance testing (NHST) on the group mean response times implies the research question is whether the theory that all condition means are equal (i.e., there is no difference in accuracy between the different conditions) can be rejected. The actual research question might deviate from this assumption in two ways. First, the researcher might not be interested in rejecting the null hypothesis, but in finding evidence for a specific theory (Klugkist et al., 2011; van de Schoot et al., 2011). Specific expectations can be tested via one sided or post hoc testing in some cases (Silvapulle & Sen, 2004). Alternatively, these expectations can be evaluated directly by formulating informative or order constrained hypotheses, see Hoijtink (2012) or Chapter 11, Vanbrabant & Rosseel (2020). Second, the researcher might not be interested in whether the average response time is equal across conditions, but whether the score for each person is equal across groups.

If researchers have specific expectations, they can formulate so-called informative hypotheses (Hoijtink, 2012; Klugkist et al., 2005). Combinations of order and equality constraints can be placed on the parameters to express an informed expectation. For example, $H_1 : \mu_{similar\ word} > \mu_{different\ word} > \mu_{non-word}$ describes the expectation that the mean reaction time in the similar word condition is larger than the mean reaction time in the different word condition, which in turn is larger than the average response time in the non-word condition. Another informative hypothesis is $H_2 : \mu_{similar\ word} > \mu_{different\ word}, \mu_{non-word}$, which describes the expectation that the average reaction time in the similar word condition is larger than both other conditions, with no expected ordering between those average reaction times. Hypotheses with order constraints ('<' and '>') are also referred to as order constrained hypotheses. Such informative hypotheses can be compared to each other by means of an F-bar test (Silvapulle & Sen, 2004; Vanbrabant & Rosseel, 2020; Vanbrabant, Schoot, & Rosseel, 2015) or with Bayes factors (Hoijtink, 2012; Klugkist et al., 2005), that are used for the method in this chapter. Bayes factors are defined in Bayes' theorem, that describes how knowledge about the relative belief in hypotheses can be updated with evidence in data:

$$\frac{P(H_1)}{P(H_2)} \times \frac{P(D|H_1)}{P(D|H_2)} = \frac{P(H_1|D)}{P(H_2|D)} \tag{4.1}$$

Equation 4.1 shows how the prior odds $\frac{P(H_1)}{P(H_2)}$, the ratio of the prior probability of H_1 and H_2 can be updated with the Bayes factor $\frac{P(D|H_1)}{P(D|H_2)}$, the relative evidence in the data for H_1 and H_2 into the posterior odds $\frac{P(H_1|D)}{P(H_2|D)}$, the relative probabilities of the hypotheses, given the data. A Bayes factor then quantifies the relative evidence in the data for two hypotheses (Kass & Raftery, 1995). Thus, $BF_{12} = 10$ means that H_1 is supported 10 times more by the data than H_2. Alternatively, $BF_{12} = .5$ means that H_1 is .5 times as much supported by the data than H_2, or in other words, H_2 is $1/.5 = 2$ times more supported than H_1. In addition to compare the evidence for a pair of hypotheses, the Bayes factor can be used to find which hypothesis from a set is most supported by the data. The computation of Bayes factors used in this chapter relies on vast literature on the topic. This will not be discussed in detail here, but the interested reader is referred to Kass & Raftery (1995). The computation of Bayes factors for informative hypotheses with inequality constraints is described in Hoijtink (2012), Klugkist et al. (2005), Klugkist, Laudy, & Hoijtink (2010) and Mulder, Hoijtink, & Klugkist (2010).

Common statistical analyses, like ANOVA and regression, test for the presence of group level effects. If $BF_{12} = 10$, we have 10 times more support that the mean reaction times are ordered like in H_1 compared to the ordering in H_2. However, if an effect detected at the group level this does not imply that the effect is true for each individual (Hamaker, 2012). For example, it might be that for part of the population H_1 reflects the true average reaction times well, but that for another part of the population, there is no effect of condition (H_0). At the group level, the conditions appear to have an effect, but this is not true for every individual. A researcher might not be interested in the average differences between groups, but in the **individual* effects (Haaf & Rouder, 2017; Molenaar, 2004). The data can also be used to analyze hypotheses on a case by case level by computing a BF for each individual. If a researcher is interested in answering the question whether an informative hypothesis holds for everyone in a sample, he needs to be able to synthesize the BFs from single-case analyses into an aggregate BF.

4.3 Data, model and hypotheses

This section introduces the Time Estimation dataset that is used as an example throughout this chapter. The individual level model and hypotheses considered for this dataset are presented. The first paragraph introduces the model at the individual level. The next paragraphs introduce the informative hypotheses considered for the parameters in this model. The difference between individual and average hypotheses is discussed.

The Time Estmation dataset is presented in Ham (2019). This dataset consists of the results of a within-subject experiment where 29 participants were each exposed to movie clips in 2 conditions. In each condition, participants watched 10 movie clips of $7 - 90$ seconds and rated the *emotional valence* and *arousal* on a 9-point Likert scale they experienced after each clip and estimated the *duration* of the clip. The content of the movies was chosen such that the set contained a range of levels of arousal and emotional valence (e.g., starving lion, coconut shells; Ham (2019)). Of the 20 movie clips in total, 10 were presented in the *Virtual Reality* (VR) condition, where participants wore a VR headset and 10 clips were presented in a *real life* (RL) scenario in the cinema. The main interest of this experiment

is the effect of *condition*, *valence* and *arousal* on the relative time estimation. That is, the interest is in the extent to which the mode of watching a clip, its perceived valence and arousal affect how much duration estimates deviate relative to the true duration.

4.3.1 Individual Level Model

Testing whether a hypothesis holds for all individuals requires data to be collected for multiple individuals and have multiple measurements for each person to estimate the individual parameters. The example data illustrated in the previous section has a nested structure. That is, the available measurements are nested within individuals. For each person $i = 1, ..., N$ a complete dataset of 20 measurements is available. Since the interest of the research is to measure the effect of *Valence*, *Arousal* and *Condition* on the *Relative Time Estimation*, the data are modelled using the following regression model:

$$\text{RelTimeEst}^i_j = \beta^i_0 + \beta^i_c \text{Condition}^i_j + \beta^i_v \text{Valence}^i_j + \beta^i_a \text{Arousal}^i_j + e^i_j \tag{4.2}$$

where the Relative Time Estimation (RelTimeEst) of person i to movie $j = 1, ..., J$ is predicted based on the Condition that movie was presented in (VR = 0, RL = 1), the rated Valence and the rated Arousal of the movie clip. By modelling the data for each individual in a separate regression model we can make predictions at the individual level.

4.3.2 Hypotheses

Hypotheses can be formed for the parameters of any individual model. A researcher could be interested in testing the null hypothesis

$$H^i_0 : \beta^i_{condition} = \beta^i_{valence} = \beta^i_{arousal} = 0 \tag{4.3}$$

Note that H^i_0 is the null hypothesis for person i, meaning that N=29 null hypotheses can be formulated. The superscript differentiates H^i_0 from the average null hypothesis

$$H_0 : \beta_{\text{condition}} = \beta_{\text{valence}} = \beta_{\text{arousal}} = 0 \tag{4.4}$$

that hypothesizes the average effect of condition, valence and arousal to be all zero.

A researcher could be interested in whether for all participants H_0ˆi is a good hypothesis. This can be represented in the following so-called For-all-hypothesis:

$$H^{\forall i}_{(\cdot)} : H^1_{(\cdot)} \& \ldots \& H^i_{(\cdot)} \& \ldots \& H^I_{(\cdot)} \tag{4.5}$$

where the superscript $\forall i$ means that for all $i = 1, \ldots, N$, the subscript $(\cdot)$ indicates a common hypothesis number such that the For-all-hypothesis $H^{\forall i}_{(\cdot)}$ expresses the expecatation that $H^i_{(\cdot)}$ holds for all individuals i.

Ham (2019) were not interested in testing the null hypothesis as shown in Equation 4.4. Rather, they had formulated three informative hypotheses about the population regression coefficients. These hypotheses are presented in the left column of Table 4.1. H_1 specifies the expectation that that there is no effect of condition, while valance and arousal have a positive effect on relative time estimation, while H_2 describes the expectation that all regression coefficients are positive. Finally, H_{1c} is the complement of H_1 and specifies the expectation that at least one of the regression coefficients for arousal and valence is not positive or that the effect of condition is different from zero. The equivalent of these average hypotheses was considered at the individual level. These individual hypotheses are presented in the right column of Table 4.1. The only difference with the population level hypotheses is that the hypotheses now concern individual regression coefficients rather than population regression coefficients.

Table 4.1
Hypotheses considered for the Time Estimation data.

Population hypotheses	Individual hypotheses
$H_1 : \beta_c = 0, \beta_v > 0, \beta_a > 0$	$H_1^i : \beta_c^i = 0, \beta_v^i > 0, \beta_a^i > 0$
$H_2 : \beta_c > 0, \beta_v > 0, \beta_a > 0$	$H_2^i : \beta_c^i > 0, \beta_v^i > 0, \beta_a^i > 0$
$H_{1c} :$ *not* H_1	$H_{1c}^i :$ not H_1^i

Note. The left column presents the population hypotheses considered in the original paper by Van der Ham et al. (2019). The right column presents the equivalence of these hypotheses in subject-specific hypotheses, considered in the current chapter. $\beta_c = \beta_{condition}$, $\beta_v = \beta_{valence}$ and $\beta_a = \beta_{arousal}$.

Summing up, this chapter considers evaluating the same hypothesis at the individual level over a group of individuals, to evaluate whether a theory holds for everyone. These hypotheses can take the form of informative hypotheses, that are translated expectations from theories, rather than a standard null or alternative hypothesis.

4.4 Individual Bayes factors

Bayesian statistics is well suited to compare multiple hypotheses, whether they are null hypotheses, unconstrained or informative, like introduced in the previous section. A Bayes factor quantifies the relative evidence in the data for two hypotheses (Kass & Raftery, 1995). More specifically, a Bayes factor is the rate with which the prior beliefs are updated into posterior beliefs, as shown in Equation 4.1. That is, prior to data collection, a researcher already has knowledge about the probability of two hypotheses, that can be quantified to express their relative probability. For example, if the researcher expects both hypotheses equally probable before observing the data, the prior ratio is 0.5/0.5. The prior ratio is updated with data and the resulting Bayes factor then quantifies how the data influenced this prior knowledge, summing up an updated ratio. Bayes factors are mostly used to evaluate hypotheses on population effects (i.e., there are no differences in the average reaction times between the conditions). In this chapter, the interest is in describing the relative evidence

for two hypotheses for a specific individual. For this purpose, the BF can be computed per subject. Section 4.5 demonstrates how this individual level evidence can be synthesized.

To analyze individual hypotheses presented in Table 4.1 using *BF*s, two steps need to be executed. First, the hypotheses need to be evaluated separately which is described in this section. In the next section it is demonstrated how the individual Bayes factors can be aggregated. For binomial data (e.g., number of successful trials per condition), a stand-alone Shiny application is also available to evaluate and aggregate individual level hypotheses (Klaassen, Zedelius, Veling, Aarts, & Hoijtink, 2017).

4.4.1 Analysis

The R (R Core Team, 2013) package `bain`, developed by (Gu, Mulder, & Hoijtink, 2017) was used to evaluate informative hypotheses for each person.

All code presented in this chapter is also available on https://github.com/fayetteklaassen/gpbf. To read the data into R the following code can be used:

```
# install bain
install.packages("bain")
# load bain
library("bain")
```

Next, the data can be loaded with:

```
# read data from the online repository
data <- read.table(file =
  "https://raw.githubusercontent.com/fayetteklaassen
  /gpbf/master/data.txt",
  header = TRUE)
# Determine the number of unique ppnrs = the number of cases
 N <- length(unique(data$ppnr))
```

Next, a Bayes factor has to be computed for each person, for the hypotheses in Table 4.1 The code below first creates an empty list to store the results of each person in. Inspecting the `names()` of the data, tells us how the variables are stored in `R`, so that these names can be used in later functions. A random seed is set to make the results replicable. Next, a loop over all subjects is created, such that the data of that subject is selected. The function `bain()` requires the estimates of the linear model as input. These are obtained by running the linear regression model `lm()`, where TimePerception is predicted by Condition, Valence and Arousal. Finally, the function `bain()` is executed, where the estimates of the linear model for person *i* are used to evaluate the hypotheses provided. The hypotheses can be entered in quotation marks, separating hypotheses by a semicolon. The names of the variables that were inspected earlier can be used to refer to the relevant regression coefficients.

```
# create an empty list to store results
 results <- vector("list", length = N)
names(data)
```

```
## [1] "ppnr"           "TimePerception" "Valence"        "Arousal"
## [5] "Condition"
```

```
set.seed(7561) # seed to create replicable results
for(i in 1:N) {  # loop over N individuals
  data_i <- data[data$ppnr == i,]   # subset data for ppnr == i

  fit_i <- lm(formula = TimePerception ~ Condition + Valence +
                                         Arousal,
              data = data_i)   # execute linear model
  # save the results of Bain analysis.
  results[[i]] <-  bain(fit_i, "Condition>0 & Valence>0 & Arousal>0;
                            Condition=0 & Valence>0 & Arousal>0")
}
```

4.4.2 Results

To obtain the final results, the code below can be executed. First, looking at the names of the bain output for the first person tells us there is an object named fit and a BFmatrix resulting from the analysis. The column labeled BF (the seventh column) of the fit object contains the Bayes factors of each hypothesis, H_1^i and H_2^i in Table 4.1, against their complement (H_{1c}^i and H_{2c}^i). The BFmatrix contains the Bayes factors comparing H_1^i to H_2^i and vice versa. The first row and second column contains the BF_{12}^i.

```
# view the names of the bain output for first person ([[1]])
names(results[[1]])
```

```
##  [1] "fit"                      "BFmatrix"
##  [3] "b"                        "prior"
##  [5] "posterior"                "call"
##  [7] "model"                    "hypotheses"
##  [9] "independent_restrictions" "estimates"
## [11] "n"
```

```
# view the output of fit and BFmatrix
results[[1]]$fit
```

```
##      Fit_eq   Com_eq      Fit_in    Com_in         Fit       Com
## H1 1.000000 1.000000 0.003379835 0.1510565 0.003379835 0.1510565
## H2 0.691829 1.627527 0.091963982 0.2379305 0.063623352 0.3872382
## Hu       NA       NA          NA        NA          NA        NA
##            BF      PMPa      PMPb
## H1 0.01905922 0.1198588 0.0188549
## H2 0.16430028 0.8801412 0.1384543
## Hu         NA        NA 0.8426908
```

```
results[[1]]$BFmatrix
```

```
##          H1        H2
## H1 1.000000 0.1361814
## H2 7.343149 1.0000000
```

To collect the relevant results for all subjects, the following code can be used. First, an output table is created, with two columns and N rows. Next, a loop over all persons saves the relevant Bayes factors in this output matrix.

```
# create output table with N rows and 4 columns
output <- matrix(0, nrow = N, ncol = 2)
# name the columns of the output
colnames(output) <- c("BF1c", "BF12")

# loop over persons
for(i in 1:N){
  # obtain the fit table of person i
  BFtab <- results[[i]]$fit
  # extract relevant BFs
  # row 1 (hypothesis 1), column 7 (BF H1 vs complement)
  BF1c <- results[[i]]$fit[1,7]
  # BF H1 vs H2
  BF12 <- results[[i]]$BFmatrix[1,2]
  # save BFs in the i-th row of the output matrix
  output[i,] <- c(BF1c,BF12)
}
# view the final output
output
```

The individual Bayes factors are presented in Table 4.2. The table shows that H_1^i is preferred over H_{1c}^i for 16 out of 29 subjects, and preferred over H_2^i for 22 out of 29 subjects. The next step is to synthesize this evidence into an aggregated BF for $H_{(\cdot)}^{\forall i}$.

4.5 Aggregating Bayes factors

Independent Bayes factors can be aggregated into a combined Bayes factor by taking their product (Klaassen et al., 2017). The interpretation of this product is the evidence that H_1 is preferred over H_2 for persons $1, ..., N$, where N is the number of individuals. This again shows that the individuals are evaluated separately: their evidence is combined but kept intact at the individual level. The scale of this product depends on the number of observations included and is therefore difficult to compare from study to study. To make the output comparable over studies, we can take the geometric mean of the product of Bayes factors, the gPBF. This is the equivalent to an average, but then for products rather than a sum. The gPBF is the average relative evidence for two hypotheses in an individual. We can evaluate how many of the individual Bayes factors describe evidence in favor of the same hypothesis as the gPBF. This is called the Evidence Rate (ER). The ER is used to evaluate to what extent individuals indeed come from the same population. If the ER is 1, all individuals show evidence for the preferred hypothesis by the gPBF. If the ER is (near) 0, almost no individuals show evidence for the preferred hypothesis by the gPBF. Another measure that can be used to evaluate the geometric mean and the individual Bayes factors is the Stability Rate (SR). This is the proportion of individual Bayes factors that expresses evidence for the same hypothesis as the gPBF, but with stronger evidence. This

quantifies the (in)balance of individual Bayes factors. If it is .5, the gPBF is affected equally by larger a smaller Bayes factors, while if it is close to 1, most cases express evidence stronger than the mean itself, and only a few cases with relatively weak or reverse evidence diminish the effect. If the SR is close to 0, this indicates that the gPBF is determined by a few strong cases, with most other cases expressing weaker evidence or reverse evidence. Together, these three measures (gPBF, ER and SR) provide information about how uniform the population can be expected to be with regard to the considered hypotheses, and what the expected relative evidence is for a next person. In what follows it is explained how the evidence of multiple individual Bayes factors can be aggregated to answer the question 'does everyone?'. The results of the example analysis are presented and interpreted.

4.5.1 Analysis

The individual Bayes factors can then be aggregated using a function available on www.github.com/fayetteklaassen/gpbf. The function requires as input a matrix with N rows and K columns, where N represents the total of individuals and K the number of Bayes factors for which the aggregate conclusion is of interest. The output of the individual analyses created in the previous section fulfills this requirement and can be used in the function. The output of the function is a list that contains: a table containing the gPBF for all Bayes factors considered; the individual Bayes factors used as input; and the sample size N.

```
# execute the gPBF function on the output from previous section
gpout <- gPBF(output)
# view the output
gpout
```

The function can be applied to any collection of individual Bayes factors. If you use your own software to compute Bayes factors at the individual level, and create a matrix of N rows and K columns, the function `gPBF()` can be applied. This function computes the geometric product over all N individuals for each of the K comparisons of interest (for example, $K = 3$, with BF_{12}, BF_{1c} and BF_{12}). The Evidence Rate is computed as the proportion of individual BFs support the same hypothesis as the gPBF, and the SR is computed as the proportion of individual BFs that express stronger evidence as the gPBF.

4.5.2 Results

Table 4.3 presents the geometric means of the product of individual Bayes factors (gPBF), the Evidence Rate (ER) and the Stability Rate (SR) for the Time Estimation data.

The results show that based on the gPBF there is no clear evidence that H_1^i is preferred over H_{1c}^i or H_2^i for everyone, or vice versa. Specifically, Table 4.3 shows that $gPBF_{1c} = .649$, indicating that the average individual evidence is 1.54 times stronger in favor of H_{1c}^i compared to H_1^i. The ER for $BF_{1c}^{\forall i}$ shows that the proportion of individual Bayes factors preferring H_{1c}^i is .448, quantifying the earlier observation that 44.8% of individual Bayes factors prefer H_{1c}^i over H_1^i. The SR of .345 indicates that there are relatively few cases expressing stronger evidence in favor of H_1^i than the gPBF. Together with the weak evidence, this indicates that the hypotheses do not describe the subjects well as a group together.

Neither the informative hypothesis, nor its complement can predict the group of subjects adequately. For the comparison H_1^i to H_2^i we find that the gPBF is 1.130, not indicating a clear preference for either hypothesis. The ER of .759 tells us that most subjects express support for H_1^i, and the SR of .690 indicates that the gPBF is influenced somewhat by strong evidence for H_2^i by some subjects. Indeed, Table 4.2 shows that subjects 5, 11 and 21 express relatively strong evidence for H_1^i (a factor of 16.67 or higher). The results indicate that the hypotheses considered are not likely to hold for all subjects. Moreover, it seems possible that while H_1^i might be a better description than H_2^i for some subjects, it does clearly not apply to all individuals.

In the group-level analysis an average preference for H_1 over both H_{1c} and H_2 was found (Ham, 2019). These analyses cannot be compared thoughtlessly. After all, in the group-level model, individual effects are shrunk to the average effect and dependent on another. However, we do get some insight that on average H_1 seems to be a good model, while it appears from the individual analysis to not hold for all individuals. Future research could develop new theories that might indeed describe all individuals, or try to explain the separation in effects found in the individual analysis. Perhaps an unmeasured variable explains why for some individuals H_1^i is preferred over H_2^i and for others not.

4.6 Conclusion and limitations

This chapter has demonstrated how one can evaluate whether a hypothesis is supported for all individuals. To answer such question, the geometric Bayes factor was introduced, which synthesizes the evidence from multiple individuals. The goal of this chapter is twofold. First, it invites researchers to rethink their own research questions and hypotheses. What is the goal of an experiment? Is it to show average effects, or demonstrate the iniquitousness of a theory? If an effect, theory or model holds on average in a population, this is no proof of the existence of such an effect in any individual specifically. Second, if indeed a researcher is interested to investigate whether a hypothesis is supported by everyone, this chapter presents the steps required to analyse this question and how to draw conclusions. The methodology is easy to use and apply to users already familiar with Bayesian (order constrained) hypothesis testing.

The data required for the proposed methodology can also be analysed with multilevel models. Multiple measurements are required for each person in each condition to be able to draw inference about individual effects. In a multilevel model this data can be modeled for example by including random effects that account for the dependency between individual subjects, in order to generalize to a population effect. By enforcing that individual effects are normally distributed around the average effect, a phenomenon called shrinking occurs: the individual effects are being pulled towards the mean, see (Chapter 5, Van Erp, 2020). A multilevel model can be used to test the variance of individual effects, but not to evaluate whether a hypothesis applies to each individual separately. The methodology in this chapter answers a different question, namely whether the evidence at the individual level is homogeneous over a sample of individuals.

It is important to keep in mind that the consistency of a Bayes factor depends on sample size. For the methodology presented in this chapter, that implies that the number of subjects

and measures per condition are both important. The number of subjects affects the stability of the ER and SR (Klaassen et al., 2017), while the number of replications affects the consistency of the individual Bayes factors. Another important consideration is the number of hypotheses to consider in a comparison. The more hypotheses are considered in a set, the more difficult it is to find one clear best hypothesis.

Table 4.2
Individual Bayes factors

Person	BF_{1c}	BF_{12}
1	0.16	7.25
2	0.96	1.74
3	2.77	1.77
4	0.17	6.97
5	0.00	0.00
6	3.25	2.83
7	1.52	2.64
8	8.10	1.57
9	5.48	0.96
10	0.70	9.55
11	0.03	0.01
12	3.66	1.79
13	0.05	8.24
14	0.27	1.02
15	3.71	2.98
16	3.02	3.40
17	0.09	0.20
18	5.39	3.33
19	0.34	4.35
20	1.08	3.36
21	0.08	0.06
22	2.37	1.04
23	1.30	2.43
24	2.50	0.27
25	0.37	6.16
26	2.30	1.87
27	0.80	0.50
28	2.18	2.22
29	1.49	3.12

Table 4.3
Aggregated Bayes factors

	BF_{1c}	BF_{12}
Geometric Product	0.649	1.130
Evidence Rate	0.448	0.759
Stability Rate	0.345	0.690

Chapter 5

Staying in the loop: Prior odds, Bayes factor, posterior odds

by F. Klaassen[1]

5.1 Introduction

Updating knowledge is a key part of scientific research. One of the first concepts discussed in any introductory methodology and statistics course is the scientific cycle (e.g. Neuman, 2011, pp. 14–18), that discusses the updating of theories. Theory and observations form the basis for new research questions and hypotheses. By collecting and analyzing data, researchers try to answer these questions. From this new state of knowledge theories can be further verified, fine-tuned or used to inform policy. Continuously going through this cycle ensures that you are 'staying in the loop'. Updating is at the foundation of Bayesian statistics, visible in Bayes' theorem:

$$P(A|D) = \frac{P(A) \times P(D|A)}{P(D)} \propto P(A) \times P(D|A) \tag{5.1}$$

that demonstrates how we can update our prior knowledge $P(A)$ about property A with data D into posterior knowledge $P(A|D)$ about A, conditional on D. Equation 5.1 is used to updated knowledge about parameters or hypotheses (substitute θ or H for A, respectively). Figure 5.1 illustrates updating at the level of theories, hypothesis probabilities and parameter distributions.

Parallel in the cycles is some state of prior knowledge that is updated with data, evidence or an answer, into posterior knowledge. The actual updating step links the cycles together.

[1]Manuscript under review at Journal of Mathematical Psychology.

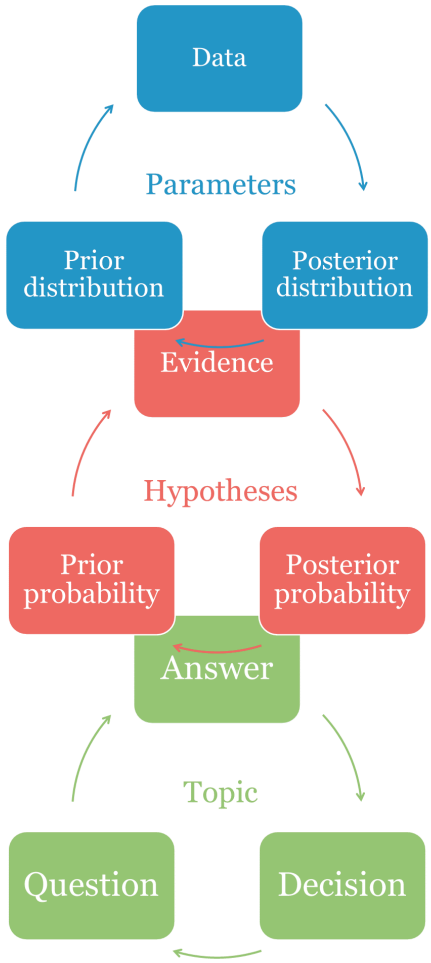


Figure 5.1. Three updating cycles. The top cycle depicts how data is used to update a prior into a posterior distribution of parameters. The middle cycle depicts how evidence obtained using the top cycle is used to update prior probabilities into posterior probabilities of hypotheses. Finally, the bottom cycle depicts how a research question can be answered and acted upon using the posterior probabilities from the middle cycle.

5.1.1 Topic

The bottom loop of Figure 5.1 illustrates updating at the level of theories. Specifically, it shows that by answering a research question new questions are generated. To illustrate this updating cycle, let us consider dr. Jones, a researcher who investigates the prevention of headaches. Currently, she is interested in the question whether a new drug is an effective headache cure. After her first research indicates that the new drug is likely effective against headaches, dr. Jones develops new research questions about the size of the effect and the side effects of the new drug. Before being able to update her research question and theories, she needs to answer her initial research question.

5.1.2 Hypotheses

The middle loop of Figure 5.1 shows that an answer can be obtained by updating knowledge about a set of hypotheses. Dr. Jones expects that a new drug performs better than paracetamol, which in turn outperforms the placebo. Alternatively she also considers the possibility that paracetamol outperforms both the placebo and the new drug. She translates these expectations into two hypotheses: H_1 : effect new drug > effect paracetamol > effect placebo and H_2 : effect paracetamol > {effect new drug, effect placebo}. Note that dr. Jones' expectations describe orderings (> and < denote larger than and smaller than, respectively) between group means rather than equalities like in a null hypothesis. The remainder of this paper is illustrated with such inequality constrained – informative – hypotheses (Hoijtink, 2012; Klugkist et al., 2005). Updating knowledge at the level of hypotheses can be illustrated by means of Bayes' theorem (Equation 5.1):

$$\frac{P(H_a)}{P(H_b)} \times \frac{P(D|H_a)}{P(D|H_b)} = \frac{P(H_a|D)}{P(H_b|D)}, \tag{5.2}$$

where $P(H_a)$ is the prior probability of H_a, $a, b = 1, 2, ..., I$, $a \neq b$ and I is the number of considered hypotheses, $P(D|H_a)$ is the marginal likelihood of data D under H_a and $P(H_a|D)$ is the posterior probability of H_a. The ratios of prior and posterior probabilities are also called the prior and posterior odds, respectively. The ratio of two marginal likelihoods is commonly called a Bayes factor (Kass & Raftery, 1995), such that BF_{ab} quantifies the relative evidence for H_a and H_b. Equation 5.2 can also be written as:

$$\text{PrO}_{ab} \times \text{BF}_{ab} = \text{PoO}_{ab}, \tag{5.3}$$

that is, the prior odds PrO_{ab} are updated with BF_{ab} – the relative belief in the two hypotheses after observing data D – into the posterior odds PoO_{ab}.

This evidence describes the rate with which the relative belief in two hypotheses changes. Dr. Jones needs to quantify her knowledge about the two hypotheses into prior probabilities to obtain the posterior odds that answer her research question. The goal of this paper is to provide a definition of what these prior probabilities are and present a procedure of how to obtain them. For now let us assume dr. Jones knows what prior probabilities to consider for her hypotheses.

5.1.3 Parameters

The top loop of Figure 5.1 shows that the evidence can be obtained by updating knowledge about a set of parameters. Dr. Jones considers parameters $\theta_{\text{paracetamol}}$, θ_{placebo} and $\theta_{\text{new drug}}$ in her hypotheses, where θ denotes the mean reduction in headache complaints in the respective groups, such that her hypotheses now are:

$$H_1 : \theta_{\text{new drug}} > \theta_{\text{paracetamol}} > \theta_{\text{placebo}} \tag{5.4}$$

and

$$H_2 : \theta_{\text{paracetamol}} > \{\theta_{\text{new drug}}, \theta_{\text{placebo}}\} \tag{5.5}$$

Bayes' theorem can again be used to show that a marginal likelihood $P(D|H)$ can be computed with:

$$P(D|H) = \frac{P(\theta|H) \times P(D|\theta, H)}{P(\theta|D, H)} \tag{5.6}$$

where $P(\theta|H)$ is the prior distribution for a set of parameters θ that quantifies the knowledge about θ before collecting any data, $P(\theta|D, H)$ is the posterior distribution of these parameters and $P(D|\theta, H)$ is the density of the data. Equation 5.2 showed how a ratio of marginal likelihoods is required to update prior odds in to posterior odds. Equation 5.6 shows that each of these marginal likelihoods depend on the prior and posterior distribution on the parameters. This demonstrated how updating the prior distributions of parameters into posteriors is required for the updating of the prior odds into posterior odds, which in turn are required to answer a research question.

Before she can update her knowledge dr. Jones needs to 1) formulate her initial theories and hypotheses, 2) specify prior probabilities and 3) define prior distributions for each hypothesis. Dr. Jones can rely on APA guidelines to help her in formulating theories and hypotheses. It is common practice to justify a research question with a literature review (e.g. VandenBos, 2010, pp. 27–28). Additionally dr. Jones' hypotheses depend on, amongst others, her background, experience, colleagues. Another researcher working in a different country, collaborating with other researchers or with more experience in the field, might develop different hypotheses. To define the prior distributions for the parameters, dr. Jones can rely on extensive literature, ranging from methodological (e.g. Gelman, Jakulin, Pittau, & Su, 2008; Mulder, 2014) to tutorials (e.g. Garthwaite, Kadane, & O'Hagan, 2005; O'Hagan et al., 2006) and reviews (e.g. O'Hagan & Perichhi, 2012).

However, guidelines for the specification of prior probabilities are scarce. While articles about updating Bayes factors discuss the importance of prior probabilities, recommendations for how to specify them lack (Mulder, 2014; Villa & Walker, 2015). Mostly no prior probabilities are considered (e.g. Hout et al., 2014; Maanen, Forstmann, Keuken, Wagenmakers, & Heathcote, 2016) or equal prior probabilities are considered under the assumption that all hypotheses are equally likely a priori (e.g. Kopp et al., 2016; Rac-Lubashevsky & Kessler, 2016). Mostly, Bayes factors are reported and interpreted as the increase in prior odds, without reporting the corresponding prior probabilities. These prior probabilities are required to complete the updating cycle at the hypotheses level, but also to update the complete set of nested updating cycles to answer the general research

question (Figure 5.1).

This paper presents a definition of prior probabilities and an elicitation procedure to specify and justify prior probabilities. The first part of this paper discusses the middle loop of Figure 5.1 in more detail. The Bayes factor is further introduced, and a definition is developed of what a prior probability is. Three meaningful components are distinguished: possibility (the probability of a hypothesis occurring, disregarding context), plausibility (the probability of a hypothesis occurring incorporating context based prior knowledge) and value (all factors that affect whether a hypothesis is considered by a researcher). Existing approaches on the specification of prior probabilities are evaluated and compared using these three concepts. The second part of this paper discusses an elicitation procedure executed with ten applied researchers. The results of this procedure demonstrate that applied researchers can define sensible prior probabilities using possibility, plausibility and value and how prior probabilities differ over persons, contexts and hypotheses.

5.2 What is a prior probability?

The goal of Bayesian hypothesis testing is to find the best hypothesis from a set of hypotheses. Consider the $j = 1, ..., J$ infinitely many potential hypotheses, of which only a subset $i = 1, ..., I$ is considered in a research project. Let us define that a prior probability $P(H_i) > 0$ quantifies that H_i is considered and that $\sum_{i=1}^{I} P(H_i) = 1$. To develop an idea of what a prior probability represents, let us consider the reasons for considering a hypothesis.

5.2.1 Possibility

A hypothesis should be possible. A hypothesis is possible (i.e., it is not impossible) if it has a probability of occurring larger than 0, disregarding the context of the hypothesis. Possibility quantifies the proportion of the parameter space a hypothesis encompasses and can take on values between 0 and 1. This definition of possibility resembles that of the complexity in the computation of Bayes factors for inequality constrained hypotheses (Klugkist et al., 2005). This is further illustrated in the next section. An example of an impossible hypothesis is $H_{\text{impossible}} : \theta_1 > \theta_2 > \theta_3 > \theta_1$. This hypothesis requires θ_1 to be larger and smaller than θ_2 and θ_3, which cannot be realized. In contrast, the unconstrained hypothesis $H_u : \theta_1, \theta_2, \theta_3$ is always true an has a probability of 1. Dr. Jones' inequality constrained hypotheses cover only a part of the parameter space.

5.2.2 Intermezzo: the Bayes factor for inequality constrained hypotheses

Klugkist et al. (2005) show that the Bayes factor for inequality constrained hypotheses can be computed without evaluating the marginal likelihoods. This approach makes use of the fact that both H_1 and H_2 are nested under the same unconstrained hypothesis $H_u : \theta_{\text{new drug}}, \theta_{\text{paracetamol}}, \theta_{\text{placebo}}$ that does not constrain the parameters in any way. If the

prior parameter distributions for the parameters in H_u are considered independent and the constrained parameters get the same diffuse prior, BF_{12} can be computed using the following equation (Klugkist et al., 2005):

$$BF_{ab} = \frac{f_a/c_a}{f_b/c_b} \tag{5.7}$$

where f_a is the fit of the data to H_a, that is, the proportion of the unconstrained posterior distribution in agreement with the constrained hypothesis H_a and c_a is the complexity of H_a, that is, the proportion of the unconstrained posterior distribution in agreement with the constrained hypothesis H_a. If the parameters in the unconstrained prior are independent (not correlated), each of the $3! = 6$ orderings of the three means is equally likely. Since H_1 is in agreement with one of these orderings, the complexity of H_1 is $\frac{1}{6}$. Two orderings agree with H_2, rendering $c_2 = \frac{1}{3}$. After collecting data, dr. Jones can determine the fit of her hypotheses and compute the Bayes factor, that can be used to obtain the posterior probabilities and answer her question.

5.2.3 Plausibility

A hypothesis should be plausible. Plausibility refers to the probability that a hypothesis occurs based on context dependent prior knowledge about the parameters. For inequality constrained hypotheses, the concept of plausibility can also be thought of as *prior fit*, that is, how well does the hypothesis fit the prior knowledge. This prior knowledge can be informed by previous experience of the effects or conditions studied or theoretical extrapolation. An example of an implausible hypothesis is $H_{\text{implausible}} : \theta_{\text{placebo}} > \theta_{\text{new drug}} > \theta_{\text{paracetamol}}$. This hypothesis is possible, but current knowledge about placebos and paracetamol tells us that this hypothesis is fairly implausible.

Dr. Jones considers her knowledge about $H_1 : \theta_{\text{new drug}} > \theta_{\text{paracetamol}} > \theta_{\text{placebo}}$ and $H_2 : \theta_{\text{paracetamol}} > \{\theta_{\text{new drug}}, \theta_{\text{placebo}}\}$. She concludes that little is known about the effectiveness of the new drug, but previous research shows that paracetamol certainly outperforms placebos. Based on this knowledge she assigns H_1 a plausibility of $1/3$. The process of how knowledge is translated into a plausibility will be the topic of discussion in Section 5.4. While assigning a plausibility to H_2, dr. Jones is unsure of the relative effectiveness of paracetamol and new drug, which combined with the knowledge of paracetamol and placebo results in a plausibility of .75.

5.2.4 Value

A hypothesis should be valued. Both possibility and plausibility contribute to how much a hypothesis is valued by a researcher. Possibility and plausibility can vary between hypotheses. A possible hypothesis can be very implausible and vice versa. For example, it is possible that the placebo will outperform both the new drug and paracetamol in preventing headaches, but it is not plausible. Alternatively, consider comparing paracetamol to not only the new drug and a placebo, but also to eating an apple, candy, a kiss on the forehead, an ice-bath and drinking a beer. It is very plausible that paracetamol outperforms all of the seven alternative 'treatments'. Each of the comparisons separately has a possibility

of .5, but combining all comparisons results in a possibility of only $.5^7 = .008$, less than 1% of the total parameter space. While this hypothesis has high plausibility, it has low possibility. The combination of possibility and plausibility in part explains why hypotheses are considered. If a hypothesis is neither possible nor plausible, it is unlikely to be considered. If a hypothesis is both possible and plausible, it might not be considered because it is considered redundant. After all, if a hypothesis has a high possibility, it is not very specific and not much can be learned from this hypothesis. If additionally, the plausibility of this hypothesis already is high, it might not be worth time and resources to learn more about this hypothesis. The hypothesis that paracetamol outperforms eating an apple or candy has low possibility and high plausibility, but is not necessarily valuable to a researcher.

Possibility and plausibility alone cannot fully explain why some hypotheses are considered, while others are excluded. Consider a researcher who wants to prove or disprove an established theory, for example: the world is flat versus the world is round. While there are many alternative hypotheses that are more possible and plausible than 'the world is flat', they do not reflect the researcher's aim and thus are not considered. The considerations for investing time and resources and thus to include a hypothesis cannot be attributed only to the possibility and plausibility of a hypothesis. Dr. Jones considers H_1 and H_2 because they are possible and plausible, and because she is particularly interested in the effectiveness of the new drug relative to paracetamol. For example, $H_{\text{not valued}} : \{\theta_{\text{paracetamol}}, \theta_{\text{new drug}}\} > \theta_{\text{placebo}}$ is both possible and plausible but dr. Jones does not consider this hypothesis because it makes no prediction on the relative effectiveness of paracetamol and the new drug.

5.3 Prior probability specification

A prior probability quantifies how much a hypothesis is valued while taking into account its possibility and plausibility. Guidelines and recommendations for specifying prior probabilities are sparse in the literature, and at best vague (Villa & Walker, 2015). The paragraphs below discuss how different approaches to prior probability specification include possibility, plausibility or value in their definition and how feasible the approach is.

5.3.1 No prior probabilities

A common practice in Bayesian hypothesis testing is to not specify any prior probabilities and focus on the Bayes factor (e.g. Wetzels, Grasman, & Wagenmakers, 2012; Hout et al., 2014). The Bayes factor can be interpreted as the strength of evidence, or as the rate with which the prior beliefs need to be adjusted (Mulder & Wagenmakers, 2016). Guidelines have been proposed to classify the strength of evidence in verbal categories (e.g. Kass & Raftery, 1995; Wagenmakers et al., 2011) that take focus away from the prior odds.

After Dr. Jones completes her research, she computes a Bayes factor and finds that H_1 is 4 *times more supported by the data than H_2. She concludes that H_1 is preferred with substantial evidence (Wagenmakers et al., 2011). She does not report any prior or posterior probabilities.*

If no prior probabilities are considered, the conclusion is only affected by the data and not by prior knowledge. This approach does not incorporate possibility, plausibility or value, because prior probabilities are not considered.

5.3.2 Equal prior probabilities

An easy way to obtain posterior probabilities is to assign equal prior probabilities to all hypotheses (e.g. Jarosz & Wiley, 2017). A simple calculation transforms any set of Bayes factors into posterior probabilities.

Dr. Jones transforms the Bayes factor using equal prior probabilities ($P(H_1) = P(H_2) = .5$. If the Bayes factor is 4, the posterior probabilities of H_1 and H_2 are .8 and .2 respectively. This tells her nothing more than what she knew already: that H_1 is 4 times more supported by the data than H_2.

Using equal prior probabilities is equivalent to interpreting only the Bayes factors (Kruschke & Liddell, 2018). Neither of these methods uses any prior information. Mulder (2014) advises to use equal prior probabilities only when absolutely no prior knowledge is available. When a researcher considers equal prior probabilities without further explanation it is unclear whether this reflects the prior knowledge or is just a default choice. Furthermore, it is very unlikely that all considered hypotheses are exactly equally probable a priori. Using equal prior probabilities is often used as a quick and default way to obtain posterior probabilities.

5.3.3 Complexity as prior probability

Jeffreys (1998, Section 1.6) elaborates the idea that simpler hypotheses should always have a larger prior probability than hypotheses that are more complex. The concept complexity was introduced in the context of inequality constrained hypotheses in the section What is a prior probability? It is a measure of how constrained a hypothesis is, relative to the unconstrained hypothesis. Using complexity as a prior probability implies that equally constrained hypotheses receive the same prior probability (Scott & Berger, 2010).

Dr. Jones considers complexity as the prior probability. She determines that $P(H_1) = 1/6$ and $P(H_2) = 1/3$ based on the proportion of the parameter space they cover.

Equation 5.7 showed that a Bayes factor for inequality constrained hypotheses can be written as a ratio of the fit of a hypothesis divided by its complexity. If the relative prior complexities are used as the prior odds and multiplied with the Bayes factor, which also contains the complexities, these cancel out and the resulting posterior odds are the relative posterior fits of the data to the model:

$$\frac{c_a}{c_b} \times \frac{f_a/c_a}{f_b/c_b} = \frac{f_a}{f_b} \tag{5.8}$$

Consequently, when comparing nested hypotheses the encompassing hypothesis will always obtain a posterior probability higher than or equal to the encompassed hypothesis. It seems

undesirable to disregard the relative specificity of a hypothesis in quantifying their relative posterior probabilities.

5.3.4 Prevalence as prior probability

Ioannidis (2005) and Wilson & Wixted (2018) suggest to use the relative prevalence rate of true hypotheses in a particular field as prior odds. Although counting the number of rejected hypotheses in a field seems like a straightforward procedure to gain knowledge about the objective plausibility of a hypothesis, three potential problems might arise. The first problem is defining the field to derive the odds from. Dr. Jones could consider all research on paracetamol, or all research on headache prevention, or all research on new versus established medication, or many other definitions of the 'field'. The second problem is that the available literature might not be a good representation of all conducted research. Publication bias creates an over-representation of significant findings in the literature (e.g. Rosenthal, 1979; Ioannidis, 2005). It is difficult to know by what factor the observed prevalence is overestimated. Finally, it is unclear how to observe the prevalence of informative hypotheses that might include combinations of (in)equality constraints and have not been considered previously.

Dr. Jones investigates the literature and finds that over $1,000$ *articles consider hypotheses similar to* H_2 *but only in about* $5\% = 50$ *of these papers is this hypothesis compared to a hypothesis like* H_1*. Only two of these studies have* H_1 *as the preferred hypothesis, and* 17 *prefer* H_2*.The remaining* 31 *papers prefer another considered hypothesis or are indecisive. Dr. Jones does not know how to translate this information into prior probabilities for her hypotheses.*

5.3.5 Subjective prior probabilities

Morey et al. (2016) and Rouder, Morey, & Wagenmakers (2016) argue in favor of defining subjective prior probabilities, that is, prior probabilities that reflect the prior beliefs of a researcher. Tijmstra (2018) also pleas for using plausibility as a reasonable prior probability. Choosing a prior probability based on the subjective beliefs of a researcher aligns with the definition of a prior probability provided in this paper. However, no methods are available for applied researchers to translate their knowledge in to a meaningful prior probability.

Dr. Jones thinks about her prior beliefs of H_1 *and* H_2*, and finds it difficult to quantify her beliefs about these hypotheses.*

5.3.6 Betting odds as prior odds

Hofstee (1984) has proposed a betting framework to think about the probabilities of hypotheses. Researchers should justify the hypotheses they choose to consider by placing a bet on the possible outcomes.

Dr. Jones considers her hypotheses and decides to bet with a rate of $6 : 2$ *on the hypotheses. From this quantification we determine that her prior probability of* H_1 *is three times larger*

than that of H_2. When she wants to report these prior odds, she has trouble explaining what made her choose these specific odds. If hypotheses are valued differently, the betting odds chosen for these hypotheses will differ. However, it is not formally described how these betting odds could be derived or defended. Similar to the consideration of subjective prior probabilities, it is unclear how a bet should be placed and how a researcher could justify the bets.

5.3.7 What is a prior probability?

The presented approaches for specifying a prior probability all seem to incorporate one or more of the reasons to consider a hypothesis (possibility, plausibility, value). If an approach used plausibility or value it is unclear how to actually quantify this. The possibility, plausibility and value together describe the prior belief in a hypothesis. The next section presents an elicitation procedure developed to elicit the possibility and plausibility for hypotheses and use this quantified knowledge to express how they value their hypotheses by betting on them.

5.4 Prior probability elicitation

A procedure was developed to elicit the prior probabilities of hypotheses. Thirteen behavioral scientists were approached to participate in an experiment in which the procedure was executed. The researchers were non-randomly selected, utilizing the network of the author within Utrecht University. A requirement for selection was familiarity with Bayesian hypothesis testing and informative (inequality constrained) hypotheses. Familiarity with Bayesian factors limits necessary explanations on Bayesian statistics, updating or prior odds. Familiarity with informative hypotheses ensures that researchers have encountered hypotheses of varying complexity before. Of the thirteen approached researchers, ten were available within the set time frame of two months (April and May 2018).

The elicitation procedure consisted of explanation and questions divided in two main parts: a Training Phase, where participants learn new concepts and procedures and a Test Phase, where the learned procedure is applied to new contexts. The whole procedure lasted approximately one hour for each participant. Informed consent about the task was obtained prior at the start of the meeting. Participants could at any point ask for clarification or guidance, and were informed that not their knowledge, but the procedure was under evaluation. Any questions or comments were answered during the elicitation by the author.

The elicitation procedure serves three goals. The first goal is to evaluate whether possibility, plausibility and value can be elicited and distinguished. This is achieved by introducing a stepwise elicitation procedure. This stepwise method is introduced to the participants after familiarizing participants with assigning probabilities to hypotheses and refreshing or extending their knowledge on Bayesian updating. The second goal of the experiment is to demonstrate that researchers are able to use the learned procedure and concepts to elicit prior probabilities for hypotheses that do not concern their own research and to their own hypotheses. The third and final goal of the experiment is to evaluate whether the procedure is a valid method to elicit prior probabilities. Throughout the elicitation

Table 5.1
Hypotheses considered at three steps in the elicitation procedure.

	No context	Headache	Flanker
H_1	$a > b\ ;\ c$	paracetamol > placebo	uniform > contrast
H_2	$a > c\ ;\ b$	paracetamol > apple	control > uniform
H_3	$a > b > c$	paracetamol > placebo > apple	uniform > control > contrast
H_4	$c > b > a$	apple > placebo > paracetamol	control > uniform > contrast

Note. The column *No context* shows the hypotheses presented to participants before context was available. The column *Headache* shows the hypotheses in the Headache example and the column *Flanker* shows the hypotheses in the flanker example.

procedure evaluation questions are asked. The results of these goals are presented in the next three sections.

5.5 Eliciting and distinguishing possibility, plausibility and value

The previous section introduced three concepts that each describe part of a prior probability: possibility, plausibility and value. The first goal of the elicitation procedure is to demonstrate that each of these concepts can be elicited. The sections below present those parts of the procedure that demonstrate the elicitation possibility, plausibility and value. The full procedure is available on https://github.com/fayetteklaassen/prior-probabilities.

5.5.1 Possibility

The possibility of a hypothesis, as discussed before, relates to the proportion of the parameter space covered by the hypothesis. To elicit this, participants are introduced to four hypotheses about parameters a, b and c. They are informed these letters represent three group means, without any further context. The hypotheses describe expected orderings between two or three of these means (see the first column of Table 5.1).

Researchers are instructed to consider each hypothesis separately and specify a probability between 0 and 1 that describes how likely it is that the hypothesis is true versus that it is not true. The possibility of a hypothesis could be derived without elicitation, but this step is explicitly introduced to facilitate the elicitation of plausibility in the next step. Participants are taught to think of possibility in terms of the number of possible orderings between parameters. Both H_1 and H_2 constrain only two parameters, while H_3 and H_4 constrain all three parameters. The assigned probabilities to the hypotheses are displayed in the first column in Figure 5.2 labeled possibility[2]. The hypotheses were chosen such that H_1 and H_2 should be equally possible and more possible than H_3 and H_4. The results show that these 10 researchers managed to quantify the possibility of four inequality constrained hypotheses without any context about these hypotheses.

[2]Note that the presented probabilities are rescaled so that their sum adds up to 1.

5.5.2 Plausibility

Plausibility of a hypothesis relates to the knowledge about the parameters. To elicit this, context is added to the example, such that participants can activate their knowledge. The same four hypotheses are considered, but now in the fictional context of a researcher who is interested in preventing headaches (middle column of Table 5.1). This randomly assigns patients to one of three treatments: a paracetamol (a), a placebo (b) or an apple (c) and measures their headache level the next day. The researchers are asked to consider the hypotheses once more and assign to each hypothesis a probability that it is true versus that it is not true, incorporating their knowledge about the headache prevention ability of paracetamol, placebo and apples. These probabilities are presented in the second column of Figure 5.2 labeled plausibility. The plausibility assigned to the hypotheses differs from the possibility of the hypotheses. Specifically, H_3 generally is considered more plausible than possible. This is reasonable because H_3 almost perfectly describes the common perceptions of the effectiveness of paracetamol, placebos and apples. Additionally, H_3 is never assigned a higher plausibility than H_1. This is a consequence of the fact that H_3 is nested in H_1, that is, H_1 has a higher possibility. Figure 5.2 shows that the plausibility assigned by researchers aligns with what can be expected based on common knowledge. It shows that for a research example where knowledge is considered fairly similar between people, the elicited plausibility indeed is rather stable over participants. Additionally, there is a clear differentiation between the assigned possibility and plausibility, where some hypotheses are assigned higher plausibility than possibility after learning the context, and for other hypotheses the reverse is observed.

5.5.3 Value

For the elicitation of possibility and plausibility researchers were asked to consider each hypothesis in itself. To evaluate whether value plays a role besides possibility and plausibility, the hypotheses are considered as a set. Three tasks in the experiment ask researchers to divide 1 euro over all hypotheses as bets, to literally measure how the hypotheses are valued. In the first task, participants are instructed to consider only the possibility they assigned to the four hypotheses before any context is added (that is, the hypotheses in the first column of Table 5.1), and consider the possibility in placing their bets. The results are presented in the column *Bet* in Figure 5.2. The placed bets are in line with what can be expected. Because H_1 and H_2 have the highest possibility, the expected pay-out for these hypotheses is highest, and there is no differentiation between the two hypotheses. Some participants choose to bet on the hypotheses with lower possibility too, but with lower bets. Finally, participant 6 bet on H_3 but not on H_4. This researcher might value H_3 and H_4 differently for how they relate to the other two hypotheses considered. This first task demonstrates that researchers already differ in how they translate the value of possibility into a bet. With nothing but the possibility to rely on H_1 and H_2 seem the most profitable to bet on. However, some people choose to bet on H_3 and H_4 as well. These hypotheses might be valued because they provide the potential knowledge gained by investigating these hypotheses.

In the second task, participants again consider the hypotheses without any context, and incorporate the knowledge gain of each hypothesis in addition to the possibility. The

knowledge gain is determined by taking the inverse of the possibility, and is presented in the form of a betting odds. That is, a hypothesis with a possibility of .5 is assigned a betting odds of $1/.5 = 2$. The betting odds tell how many times a bet is paid out if that hypothesis is in fact the best hypothesis and are a quantification of how much knowledge is gained by learning about this hypothesis. Participants are asked to consider these betting odds and the possibility they assigned to the hypotheses in placing their bets. If only the possibility of a hypothesis (transformed into the betting odds) affects how researchers bet on a set of hypotheses, the relative bets for the four hypotheses would be all equal. Consider .5 and .25 as the possibilities of two hypotheses. Their respective betting odds would be 2 and 4. The expected pay-out for each hypothesis is $.5 \times 2 = 1$ and $.25 \times 4 = 1$, which corresponds to an expected equal bet on the two hypotheses. The fourth column of Figure 5.2 shows the bets placed on the four hypotheses. While participants 1, 4, 5 and 7 indeed distribute their bets equally, the bets of participants 2, 6 and 10 resemble their assigned possibilities. In other words, these participants only considered the possibility in placing their bets. Participants 3 and 8 seemed to consider only the betting odds in placing their bets, betting more on the hypotheses that pay out more (provide more knowledge). Finally, participant 9 considered even a different approach, betting most on H_2 and H_4. While there is no numerical reason to make this particular bet, it appears this participant considers something more than the possibility and the knowledge gain of the hypotheses. The reason might be that the comparison of H_2 and H_4 seems the most interesting to this researcher. This second task demonstrates that researchers show different betting behavior that cannot uniquely be attributed to possibility or the gain in knowledge.

Plausibility too can play a role determining the prior probabilities of hypotheses. The third task asks researchers again to place a bet on the hypotheses, taking into account the betting odds, plausibility and how they value the hypotheses. Similar to the previous task, the expected pay-off can be computed, by multiplying the betting odds (the pay-out) with the plausibility (the subjective probability that the hypothesis is true). The expected value is presented in the fifth column of Figure 5.2 labeled *Predicted bet*. The sixth column, labeled *Bet 3* presents the actual placed bets. For participants 2, 5, 6, 7 and 8, the relative ordering of placed bets resembles the prediction. Only the size of the bets deviates from the prediction. The bet placed on H_4 is higher than expected for participants 4 and 9, indicating that this hypothesis is valued more than expressed by only the possibility and plausibility. Even though the participants have similar knowledge on the effectiveness of paracetamol, placebo and an apple in preventing headaches, the final bets are widely different from each other and from the expectation if only possibility and plausibility are taken into account. It appears that the hypotheses are valued differently by different researchers.

The Training Phase shows three things. First, it shows that ten researchers were able to specify the possibility of four hypotheses. Second, after adding a context to the hypotheses researchers specify the plausibility of hypotheses. The added context affects the variability of the individual answers, indicating that the plausibility of hypotheses differs from person to person, albeit slightly. Finally, when asked to placed bets incorporating only possibility or both possibility and plausibility, the placed bets differ from expected bets, indicating that individuals value the hypotheses differently. The final bet placed is an elicited prior probability. In this prior probability, researchers are given the opportunity to include the possibility, plausibility and their value of the hypotheses.

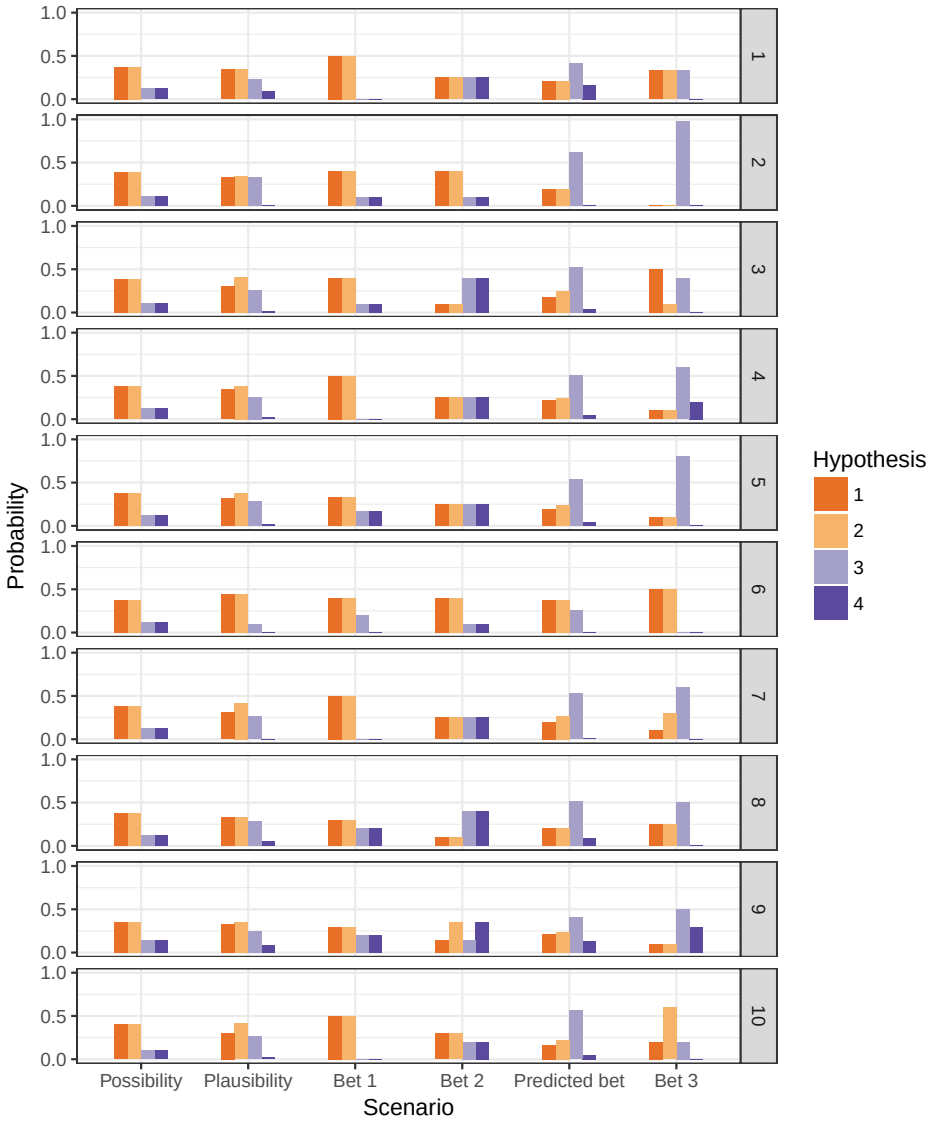


Figure 5.2. Assignment of possibility, plausibility and bets to the hypotheses without and with context for the headache example (see Table 5.1). Each row depicts a participant. The columns show (1) Possibility; (2) Plausibility; (3) Bet 1, incorporating possibility; (4) Bet 2, incorporating possibility and betting odds; (5) Prediction, the predicted bet based on betting odds and plausibility; (6) Bet 3, incorporating betting odds and plausibility.

5.6 Eliciting probabilities

The second goal of the elicitation procedure is to evaluate whether researchers can use the procedure to assign probabilities in practice. This is achieved by two final tasks in the experiment. In the first task participants are asked to consider four hypotheses about a psychological phenomenon, that is not their own research interest. This task mimics the scenario where researchers are confronted with hypotheses and evidence of another researcher and have to define their own probabilities to evaluate this evidence. The second task asks researchers to formulate two or more hypotheses about their own research. They complete the procedure for this set of hypotheses, resulting in their own prior probabilities.

The hypotheses presented by a hypothetical other researcher concern an adaptation of the flanker task (Eriksen & Eriksen, 1974). In this fictional reaction time experiment, participants are asked to identify the middle letter in a sequence of five letters as an X (press left hand key) or an O (press right hand key) as quickly as possible. Three conditions are considered: Uniform – the target letter is flanked by copies of the same letter (e.g. XXXXX), Contrast – the target letter is flanked by copies of the contrasting target letter (e.g. XXOXX) and Control – the target letter is flanked by copies of a lower case letter that is different from the target letters (e.g. ssOss). Four hypotheses are formulated about the average reaction time in the correct trials for each of these conditions. The four hypotheses are presented in Table 5.1.

Participants are asked to think about the probability of these hypotheses, disregarding context (possibility), the probability of these hypotheses incorporating their knowledge (plausibility) and to finally place a bet on these hypotheses. Figure 5.3 shows the results of this procedure. The first column shows the possibility, which was given and is thus uniform over the hypotheses. The second column presents the plausibility of the hypotheses for each participants. The third column shows the predicted bet based on the possibility and plausibility. Finally, the fourth column shows the actually placed bets. For some participants the expected value does well in predicting the placed bets. For others, the prediction does not account for the fact that some hypotheses are not considered, by betting nothing on them. Neither the expected value nor the possibility or plausibility can explain the betting of the participants, indicating that the final probability also reflects how researchers value the hypotheses differently.

Finally, participants are asked to consider a set of hypotheses about their own research. They quantify the possibility and plausibility of their bets and place their bets accordingly. Figure 5.4 shows, from left to right, the possibility, plausibility, predicted and placed bets. Note that each researcher could define their own hypotheses, so the number and possibility of the hypotheses differ between participants. In the flanker example a bet could be withheld to express a hypothesis not valued. In the current task, researchers already excluded all hypotheses they did not value to create their set of interest, and logically all hypotheses considered receive a bet. Similar to the flanker task, the relative size of the placed bets differs from the prediction, plausibility and possibility. Even for a self-chosen set of hypotheses in a researcher's own field, plausibility and possibility cannot account for the differences in the placed bets. Additionally, between researchers different types of bets can be observed. While for participant 1, 5 and 7 the final bets are equal or almost equal for all hypotheses, participants 2, 4, 8 and 10 show a substantial bet on one hypothesis, while dividing a smaller amount over the remaining hypotheses. These differences demonstrate

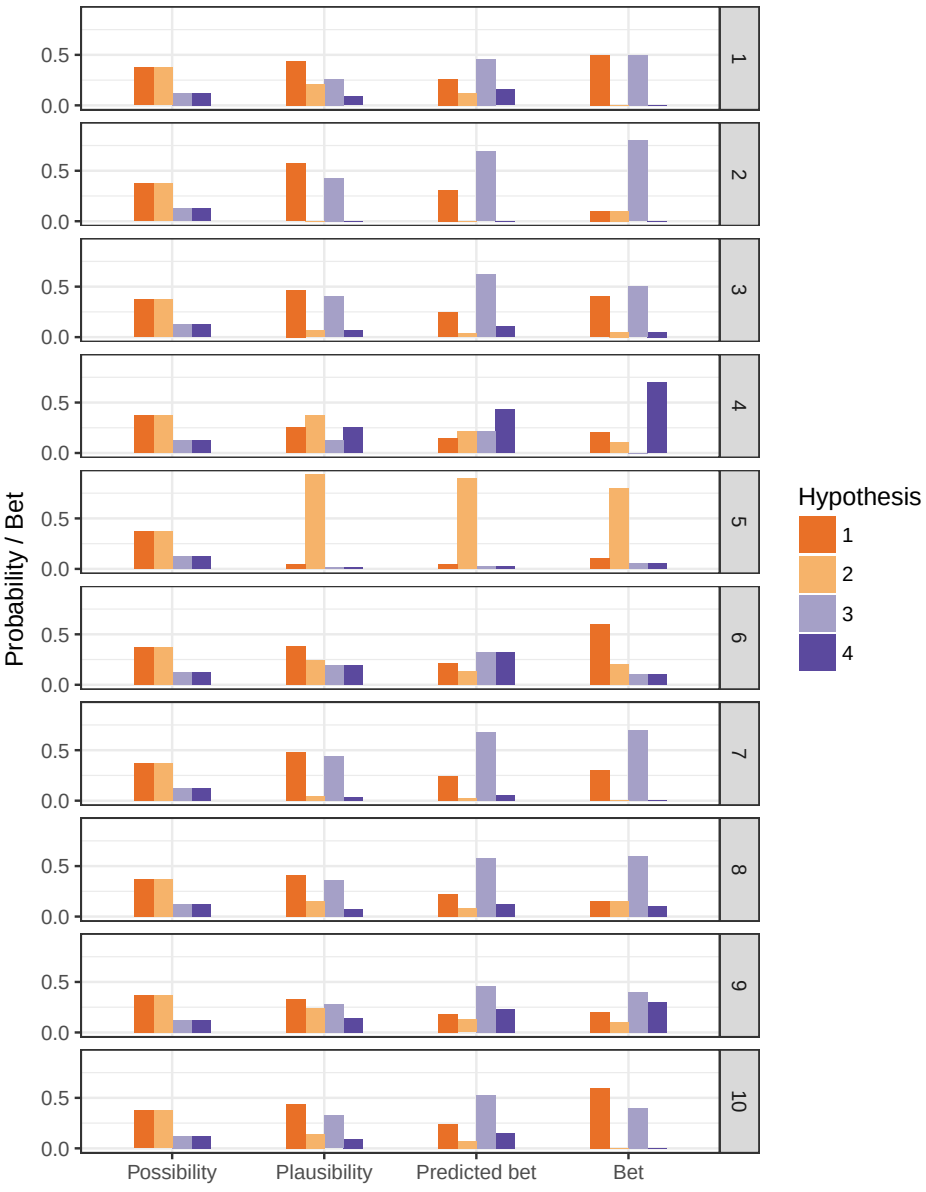


Figure 5.3. Plausibility and bets assigned to the hypotheses for the flanker example (see Table 5.1). The columns show from left to right (1) Possibility; (2) Plausibility; (3) Predicted bet based on possibility and plausibility; and (4) actually placed bet.

that it seems unrealistic to develop a default rule for prior probabilities, or a rule of thumb. Clearly, as the context of a research question differs, the set of hypotheses changes, and so do the relative probabilities.

5.7 Evaluation of the elicitation

The results of the elicitation presented above demonstrate that possibility and plausibility alone do not account for the prior probabilities of hypotheses. The value of a hypothesis is also included in the placed bets. To evaluate the elicitation method, several feedback questions were asked during the elicitation procedure. The procedure is evaluated in terms of face validity, reliability and feasibility (Johnson, Tomlinson, Hawker, Granton, & Feldman, 2010). All measures are evaluated by means of self-evaluation on a 5-point scale where 1 indicates *Not at all* and 5 indicates *Completely*. The main results of these questions are presented below.

Face validity is evaluated for each set of hypotheses (headache; flanker; own hypotheses) after placing the final bets. Participants were asked to rate how well the bets they placed reflect their knowledge for the probabilities of the hypotheses. By asking this question, researchers are invited to reflect on the input they provided so far. It appears that researchers feel mostly that the bets they specified are a good reflection of their ideas, for the probabilities of their own hypotheses (min = 4, max = 5, mean = 4.65). The average feeling of how good ideas were reflected in the bets was also good for the headache example (min = 2, max = 5, mean = 4.00) and flanker task (min = 2, max = 5, mean = 4.15). This shows that the procedure is capable of eliciting values that reflect a researcher's ideas. It appears more valid for the elicitation of probabilities for own hypotheses than for a mock example or research outside one's field. In other words, the procedure has the highest face validity when determining the prior probabilities for one's own research.

Reliability too is measured after every placed bet, by asking participants to rate the following statement on the same five point Likert scale: *If I were to complete this procedure again, I would obtain similar probabilities*. For hypotheses they specified themselves, researchers were on average quite certain about their bets (mean = 4.6, min = 4, max = 5). For hypotheses about the flanker task, participants were less certain about their own reliability (mean = 4.25, min = 4, max = 5). The self-reported reliability is high for both the prior probability of own hypotheses and for the flanker example. Even though it is self-reported, these results indicate the method is considered to provide reliable prior probabilities.

Finally, feasibility was measured at three moments in the experiment. First, at the end of the Training Phase, participants are asked whether the steps were easy to execute (mean = 4.2, min = 2, max =5) and whether they feel capable to apply the used procedure (mean = 4, min = 3, max = 5). These results show that the method was easy to follow and execute, with the provided guidance. After assigning probabilities to the flanker hypotheses, participants are asked to reflect how capable they feel in assigning probabilities to hypotheses outside of their own field (mean = 3.3, min = 2, max = 5). Finally, after assigning probabilities to their own hypotheses participants are asked to reflect how capable they feel to assign probabilities to their own hypotheses (mean = 4.05, min =4, max = 5). Similar to the face validity, participants feel more capable to assign probabilities to their own hypotheses than

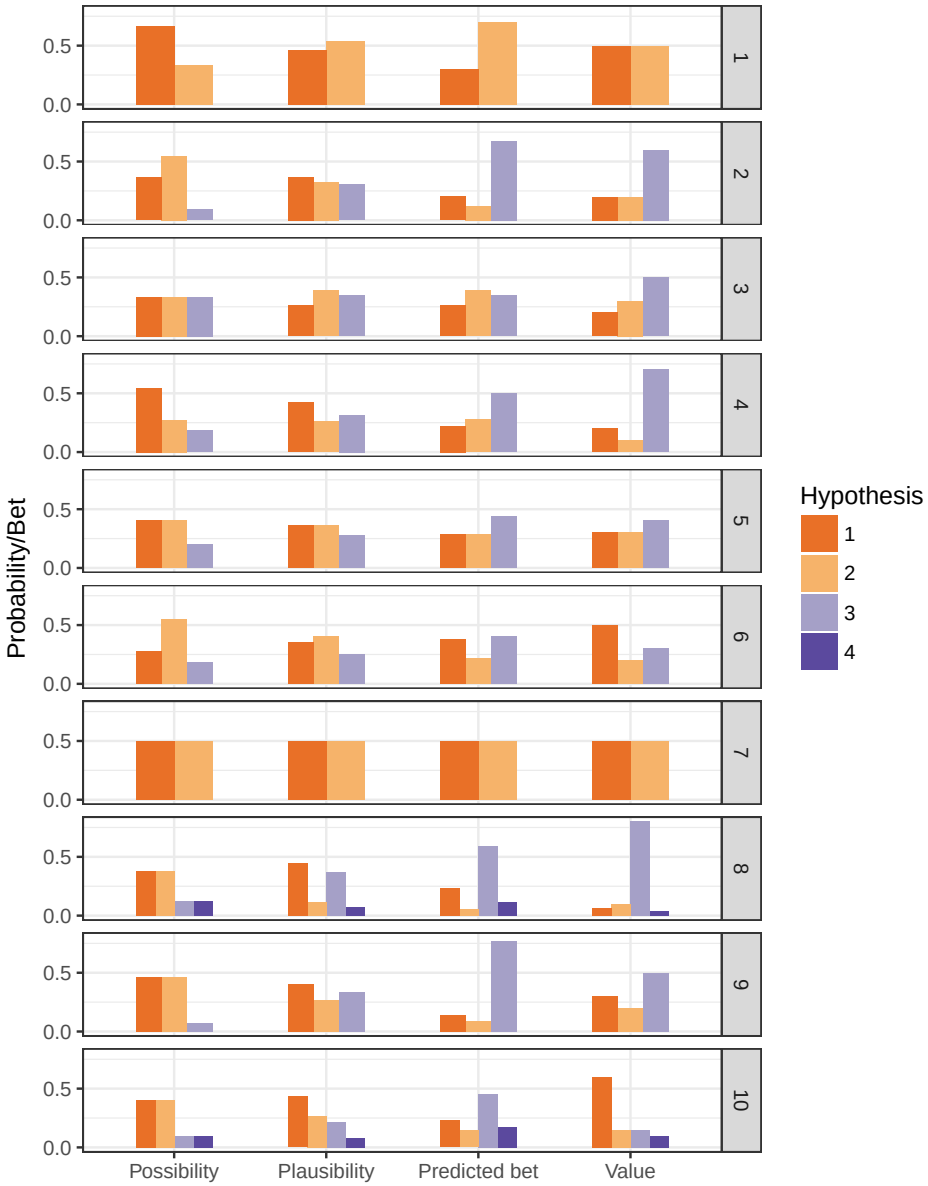


Figure 5.4. Plausibility and bets assigned to a set of hypotheses defined by each participant. The number of hypotheses varies over participants. The columns show from left to right (1) Possibility; (2) Plausibility; (3) Predicted bet based on possibility and plausibility; and (4) actually placed bet.

to the flanker hypotheses. The method might be most suitable for defining probabilities to hypotheses that a researcher has knowledge about.

5.8 Discussion and conclusion

This article has three goals. First, by discussing the role of prior probabilities within the updating cycle, the importance and necessity of specifying prior probabilities is illustrated. Posterior probabilities can only be obtained when prior probabilities are specified and updated with evidence. Dr. Jones has completed the prior probability elicitation procedure, resulting in prior probabilities of .3 and .7 for $H_1 : \theta_{\text{new drug}} > \theta_{\text{paracetamol}} > \theta_{\text{placebo}}$ and $H_2 : \theta_{\text{paracetamol}} > \{\theta_{\text{new drug}}, \theta_{\text{placebo}}\}$, respectively. She conducts an experiment resulting in a Bayes factor $BF_{12} = 4$ and update her prior odds of $.3/.7 = 0.43$ into posterior odds of 1.71. The a priori belief of dr. Jones for H_2 was almost twice as large as for H_1. The data do not align with this prior belief, and express a preference for H_1, resulting in a posterior probability of .63 for H_1 and .37 for H_2. Dr. Jones uses the posterior probabilities to answer her research question and has completed one loop of the research cycle.

However, she is not satisfied with the current state of knowledge on the effectiveness of the new drug. Based on the results of the first research project she has become interested in a third hypothesis $H_3 : \{\theta_{\text{paracetamol}}, \theta_{\text{new drug}}\} > \theta_{\text{placebo}}$. Figure 5.1 shows how the answer to a research question influences the decision to do further research, and develop a new research question. This initiates a new cycle through the updating loops. In this second cycle, the posterior odds for H_1 relative to H_2 can be used as prior odds. Because H_3 is now considered in addition to H_1 and H_2, the prior odds relative to the newly considered hypothesis have to be formulated. Also for updating the parameter cycle dr. Jones can use the posterior distribution from her first project as her new prior distribution for H_1 and H_2. For H_3, a new prior distribution has to be defined, that can be used to compute the marginal likelihood and corresponding Bayes factors. If dr. Jones had decided to not define any prior probabilities in her first research project, she would not have been able to update this knowledge in a second research cycle, because she would fail to account for the knowledge already available from the earlier research.

The second goal of this article is to define what a prior probability is. Assigning a prior probability to a hypothesis means to consider it, and thus to exclude the hypotheses that do not get a prior probability assigned. Prior probabilities can be used to describe the relative prior belief for a set of hypotheses. Three reasons determine whether a hypothesis is considered: its possibility, plausibility and value to the researcher. Existing guidelines for the specification of prior probabilities all relate to one or more of these aspects. The varying proposed strategies can be differentiated and summarized with these three components by means of the elicitation procedure in this paper. Possibility, plausibility and value can each be distinguished from the results. This seems to justify the definition of prior probabilities in terms of possibility, plausibility and value.

The third and final goal of this article is to provide a concrete guideline that allows researchers to define their own prior probabilities. This is useful in updating knowledge in their own research, where they can then update their prior probabilities and can report the findings. It can also be beneficial to interpret Bayes factors reported in other research

based on own specified prior probabilities for the hypotheses considered. This article presents an elicitation procedure that is directed at specifying prior probabilities following the distinction in possibility, plausibility and value. The results show that it is possible to differentiate these three concepts between individuals, and that by manipulating context, knowledge or value, the corresponding component is affected. Furthermore, individuals differ in their assigned prior probabilities, and rarely do equal prior probabilities represent the researchers' prior knowledge adequately.

The results also show that participants rated the procedure feasible, valid and reliable. Both the validity and self-reported reliability were on average higher in the context of own hypotheses rather than predefined hypotheses. This further supports the argument that it is important to report prior probabilities rather than leave it to the reader to define their own prior probabilities. A reader can still disagree with these probabilities, but can better consider how to disagree.

This article has demonstrated how possibility and plausibility play a role in eliciting prior probabilities, but cannot explain these entirely. How a researcher values a hypothesis relative to other considered hypotheses appears to play a role in the final prior probability assigned to it. Through evaluation of the elicitation procedure, it appears that researchers are able to think about their own prior probabilities and report their rationale behind them, whether they use the introduced procedure or another method. Specifying prior probabilities allows researchers to evaluate the evidence in light of their prior beliefs and reflect on the agreements, disagreements and surprises between prior beliefs and evidence from data. This creates a more open research cycle, and enables us to be in the loop continuously.

Chapter 6

Software

Two pieces of software have been developed alongside the research presented in this dissertation. The R package BayesianPower is published and available on CRAN [1]. Section 6.1 presents the vignette available for this package. The R Shiny application OneForAll has been developed alongside Chapter 3. This application is available online at https://utrecht-university.shinyapps.io/OneForAll/, or can be installed locally to enable simulation features. The manual of the local shiny application is presented in Section 6.2.

6.1 BayesianPower: Sample size and power for comparing inequality constrained hypotheses

BayesianPower can be used for sample size determination (using `bayes_sampsize`) and power calculation (using `bayes_power`) when Bayes factors are used to compare an inequality constrained hypothesis H_i to its complement H_c, another inequality constrained hypothesis H_j or the unconstrained hypothesis H_u. Power is defined as a combination of controlled error probabilities. The unconditional or conditional error probabilities can be controlled. Four approaches to control these probabilities are available in the methods of this package. Users are advised to read this vignette and the paper available at https://doi.org/10.17605/OSF.IO/D9EAJ where the four available approaches are presented in detail (Klaassen, Hoijtink, & Gu, n.d.).

6.1.1 Power calculation with `bayes_power()`

```
bayes_power(n, h1, h2, m1, m2,   ngroup = NULL, comp = NULL,   bound1
= 1, bound2 = 1/bound1,   datasets = 1000, nsamp = 1000,   seed =
NULL)
```

[1] Klaassen, F. (2019). *BayesianPower: Sample size and power for comparing inequality constrained hypotheses.* R packages, version 0.1.6. https://cran.r-project.org/web/packages/BayesianPower/index.html

Arguments

`n` A number. The sample size for which the error probabilities must be computed.

`h1` A constraint matrix defining H1, see below for more details.

`h2` A constraint matrix defining H2, or a character `'u'` or `'c'` for the unconstrained or complement hypothesis.

`m1` A vector of expected population means under H1 (standardized), see below for more details.

`m2` A vector of expected populations means under H2 (standardized). `m2` must be of same length as `m1`.

`ngroup` A number or `NULL` . The number of groups. If `NULL` the number of groups is determined from the length of `m1`.

`comp` A vector or `NULL` . The complexity of H1 and H2. If `NULL` the complexity is estimated. See below for more details.

`bound1` A number. The boundary above which BF12 favors H1, see below for more details.

`bound2` A number. The boundary below which BF12 favors H2.

`datasets` A number. The number of datasets to simulate to compute the error probabilities

`nsamp` A number. The number of prior or posterior samples to determine the complexity or fit.

`seed` A number. The random seed to be set.

Details

Specifying hypotheses

Hypotheses are defined by means of a constraint matrix, that specifies the ordered constraints between the means $\boldsymbol{\mu}$ using a constraint matrix R, such that $R\boldsymbol{\mu} > \mathbf{0}$, where R is a matrix with J columns and K rows, where J is the number of group means and K is the number of constraints between the means, $\boldsymbol{\mu}$ is a vector of J means and $\mathbf{0}$ is a vector of K zeros. The constraint matrix R contains a set of linear inequality constraints.

Consider

```
R <- matrix(c(1,-1,0,0,1,-1), nrow = 2, byrow = TRUE)
mu <- c(.4, .2, 0)
```

```
R
```

```
##      [,1] [,2] [,3]
## [1,]    1   -1    0
## [2,]    0    1   -1
```

```
mu
```

```
## [1] 0.4 0.2 0.0
```

```
R %*% mu
```

```
##      [,1]
## [1,]  0.2
## [2,]  0.2
```

```
(R %*% mu) > 0
```

```
##      [,1]
## [1,] TRUE
## [2,] TRUE
```

The matrix R specifies that the sum of $1 \times \mu_1$ and $-1 \times \mu_2$ and $0 \times \mu_3$ is larger than 0, **and** the sum of $0 \times \mu_1$ and $1 \times \mu_2$ and $-1 \times \mu_3$ is larger than 0. This can also be written as: $\mu_1 > \mu_2 > \mu_3$. For more information about the specification of constraint matrices, see for example (Hoijtink, 2012).

The argument `h1` has to be a constraint matrix as specified above. The argument `h2` can be either a constraint matrix, or the character `'u'` or `'c'` if the goal is to compare H_1 with H_u, the unconstrained hypothesis, or H_c the complement hypothesis.

Specifying population means

Hypothesized population means have to be defined under H_1 and H_2, also if H_u or H_c are considered as H_2. The population means have to be standardized.

Computing complexity

If the complexity of a hypothesis is known it can be specified under `comp` to reduce computational time. If `comp = NULL` the complexity is sampled using $\mu_. \sim \mathcal{N}(0, 1000)$ as a prior distribution for each mean, that is, a normal distribution with mean 0 and standard deviation 1000.

Setting bounds

`bound1` and `bound2` describe the boundary used for interpreting a Bayes factor. If `bound1 = 1`, all $BF_{12} > 1$ are considered to express evidence in favor of H_1, if `bound1 = 3`, all $BF_{12} > 3$ are considered to express evidence in favor of H_1. Similarly, `bound2` is the boundary *below* which BF_{12} is considered to express evidence in favor of H_2.

Examples

Example 1. H_1 vs H_c

An example where three group means are ordered in $H_1 : \mu_1 > \mu_2 > \mu_3$ which is compared to its complement. The power is determined for $n = 40$

```
h1 <- matrix(c(1,-1,0,0,1,-1), nrow= 2, byrow= TRUE)
h2 <- 'c'
m1 <- c(.4,.2,0)
m2 <- c(.2,0,.1)
bayes_power(40, m1, m2, h1, h2)
```

Example 2. H1 vs H2

An example where four group means are ordered in $H_1 : \mu_1 > \mu_2 > \mu_3 > \mu_4$ and in $H_2 : \mu_3 > \mu_2 > mu_4 > \mu_1$. Only Bayes factors larger than 3 are considered evidence in favor of H_1 and only Bayes factors smaller than $1/3$ are considered evidence in favor of H_2.

```
h1 <- matrix(c(1,-1,0,0,0,1,-1,0,0,0,1,-1), nrow= 3, byrow= TRUE)
h2 <- matrix(c(0,-1,1,0,0,1,0,-1,-1,0,0,1), nrow = 3, byrow= TRUE)
m1 <- c(.7,.3,.1,0)
m2 <- c(0,.4,.5,.1)
bayes_power(34, h1, h2, m1, m2, bound1 = 3, bound2 = 1/3)
```

6.1.2 Sample size determination with `bayes_sampsize()`

```
bayes_sampsize(m1, m2, h1, h2,   type = 1, cutoff, bound1 = 1,
bound2 = 1 / bound1,   datasets = 1000, nsamp = 1000,   minss =
2, maxss = 1000, seed = 31)
```

Arguments

The arguments are the same as for `bayes_power()` with the addition of:

`type`A character. The type of error to be controlled. The options are: `"1", "2", "de", "aoi", "med.1", "med.2"`. See below for more details.

`cutoff` A number. The cutoff criterion for type. If `type` is `"1", "2", "de", "aoi"`, `cutoff` must be between 0 and 1. If `type` is `"med.1"` or `"med.2"`, `cutoff` must be larger than 1. See below for more details.

`minss` A number. The minimum sample size.

`maxss` A number. The maximum sample size.

Details

`bayes_sampsize()` iteratively uses `bayes_power()` to determine the error probabilities for a sample size, evaluates whether the chosen error is below the cutoff, and adjusts the sample size.

`type`

Klaassen et al. (n.d.) describes in detail the different types of controlling error probabilities that can be considered. Specifying `"1"` or `"2"` indicates that the Type 1 or Type 2 error probability has to be controlled, respectively the probability of concluding H_2 is the best hypothesis when H_1 is true or concluding that H_1 is the best hypothesis when H_2 is true. Note that when H_1 or H_2 is considered the best hypothesis depends on the values chosen for `bound1` and `bound2`. Specifying `"de"` or `"aoi"` indicates that the Decision error probability (average of Type 1 and Type 2) or the probability of Indecision has to be controlled. Finally, specifying `" med.1"` or `"med.2"` indicates the minimum desired median BF_{12} when H_1 is true, or the minimum desired median BF_{21} when H_2 is true.

Examples

Example 1. H_1 **versus** H_c**, controlling decision error**

```
h1 <- matrix(c(1, -1, 0,
               0, 1, -1),
             nrow= 2, byrow= TRUE)
h2 <- 'c'
m1 <- c(.4, .2, 0)
m2 <- c(.2, 0, .1)
bayes_sampsize(h1, h2, m1, m2, type = "de", cutoff = .125)
```

Example 2. H1 versus H2, controlling indecision error

```
h1 <- matrix(c(1, -1, 0, 0,
               0, 1, -1, 0,
               0, 0, 1, -1),
             nrow= 3, byrow= TRUE)
h2 <- matrix(c(0, -1, 1, 0,
               0, 1, 0, -1,
               -1, 0, 0, 1),
             nrow = 3, byrow= TRUE)
m1 <- c(.7, .3, .1, 0)
m2 <- c(0, .4, .5, .1)
bayes_sampsize(h1, h2, m1, m2, type = "aoi", cutoff = .2,
               minss = 2, maxss = 500)
```

Example 3. H_1 **versus** H_u**, controlling median Bayes factor**

```
h1 <- matrix(c(1, -1, 0, 0,
               0, 1, -1, 0,
               0, 0, 1, -1),
             nrow= 3, byrow= TRUE)
h2 <- 'u'
m1 <- c(.3, .2, 0)
m2 <- c(0, 0, 0)
bayes_sampsize(h1, h2, m1, m2, type = "med.1", cutoff = 3,
```

```
        minss = 2, maxss = 500)
```

6.2 OneForAll: Multiple N = 1 Bayes factors

This manual describes how the Shiny Application 'OneForAll' can be used. A stable link to the app can be found on http://github.com/fayetteklaassen/OneForAll. The application can be run on any computer with an internet connection. By using the application, you agree to the Terms of Usage, as displayed on the starting screen of the app. This application allows you to evaluate informative hypotheses for multiple $N = 1$ studies of your own data. If you want to execute a simulation study (like presented in the paper), please contact the author at mailto for R code or a Shiny application you can run locally on your own computer.

6.2.1 Analyze own data

This section describes each of the steps required to analyze own data in the tab *Analyze own data* within the Shiny Application OneForAll. This item consists of three options from the menu: *Settings and load data*, *Individual Bayes factors*, and *GPBF output*. The first will be discussed in detail, while the other two can be used to view the results.

Step 1: Data and hypotheses

Step 1 is to select the data file to be used for analysis. You can choose to use the example data from Zedelius et al. (2011) (as described in Klaassen et al. (2017)). Alternatively, you can upload your own data file to be analyzed. This file should be a *.txt* with as many rows as persons or cases, and per row the entries for each condition, separated by a space or tab (white space). Each entry in the file should be an integer, describing the number of successes in each condition. The rows and columns should not be numbered or labeled. SPSS data can be saved as a Tab delimited .dat file (without row and column names), and the .dat extension must be manually changed to .txt . When the file is selected, a preview of the data is visible, together with a description of the number of conditions and the number of participants. If these numbers are correct, you can continue. If not, the data file was not in the right format. Common problems are that the first row contains column names (you can just delete this row), or strange symbols in the first entry, which can also be deleted. Next, the number of replications used in the experiment should be given.

Step 2: Number of conditions and hypotheses

Step 2 is to define the constraints of the hypotheses considered. Three options are available. Below examples are provided on how to use this option, using the hypotheses specified in Table 6.1.

- *Option 1: Using >.* This option requires that for each hypothesis you want to consider, you specify each constraint using > and separate constraints with a comma. Each hypothesis is specified on its own line. The parameters of interest

Table 6.1
Possible specifications for 6 hypotheses

Hypothesis	Using >	Using *R*	Default
$H_1^i : \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i > \pi_5^i > \pi_6^i$	✓	✓	✓
$H_2^i : \pi_1^i + \pi_2^i > \pi_3^i + \pi_4^i > \pi_5^i + \pi_6^i$	✓	✓	✓
$H_3^i : \pi_1^i + \pi_2^i + \pi_3^i > \pi_4^i + \pi_5^i + \pi_6^i$	x	✓	x
$H_4^i : \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i > \pi_6^i > \pi_5^i$	✓	✓	✓
$H_5^i : \pi_1^i > \pi_3^i > \pi_2^i > \pi_4^i > \pi_6^i > \pi_5^i$	✓	✓	x

are the success probabilities in the experimental conditions. They can be referred to by a number that corresponds to the column number of that condition in the data. Two types of constraints can be specified: a constraint between two parameters (e.g.1$>$2"), or a constraint between two combinations of two parameters, separated by a $'+'$ (e.g.1+2$>$3+4"). Note that one parameter cannot be on both sides of the constraint (e.g. "1+2 > 2+3" is not allowed). Figure 6.1 specifies all hypotheses from Table 6.1 that can be specified using this option.

- *Option 2: Using constraint matrix.* This options allows the user to specify a constraint matrix for each hypothesis. For more details on a constraint matrix, see Mulder, Hoijtink, & Leeuw (2012) for example. The first line should specify how many hypotheses M are specified, and each hypothesis $m = 1, ..., M$ should start with a line specifying the number of constraints (rows) in R_m. Each constraint matrix contains $J + 1$ columns, where J is the number of conditions. The first J columns specify the constraint matrix, and the last additional column should contain the contrast vector r. With this option more complex hypotheses can be specified. Figure 6.2 shows how H_1^i and H_3^i could be specified using *R*, Option 2. Option 2 is more flexible than Option 1, but as can be seen in Figures 6.1 and 6.2, Option 1 is more straightforward to specify, if the hypotheses allow for this option.

- *Option 3: Default.* This option is only available for an even number of conditions, and specifies automatically three hypotheses: $H_1^i : \pi_1^i > \pi_2^i > ... > \pi_J^i$, a full ordered hypothesis, where $J \geq 4$, $H_2^i : \pi_1^i > \pi_2^i > ... > \pi_J^i > \pi_{J-1}^i$, that only deviates from H_1^i because the last two parameters are reversed in the ordering, and finally $H_3^i : \pi_1^i + \pi_2^i > ... > \pi_{J-1}^i + \pi_J^i$, a full ordered hypothesis of each adjoining pair of parameters.

When the constraints are submitted, the third step is to specify which Bayes factors should be computed. The options available are all combinations of the hypotheses specified, and each hypothesis against its complement and the unconstrained hypothesis. By pressing the button 'Check constraints' the constraints are checked, and a textbox is returned with the hypothesis as processed by the app. If something is incorrect here, please re-enter your constraints.

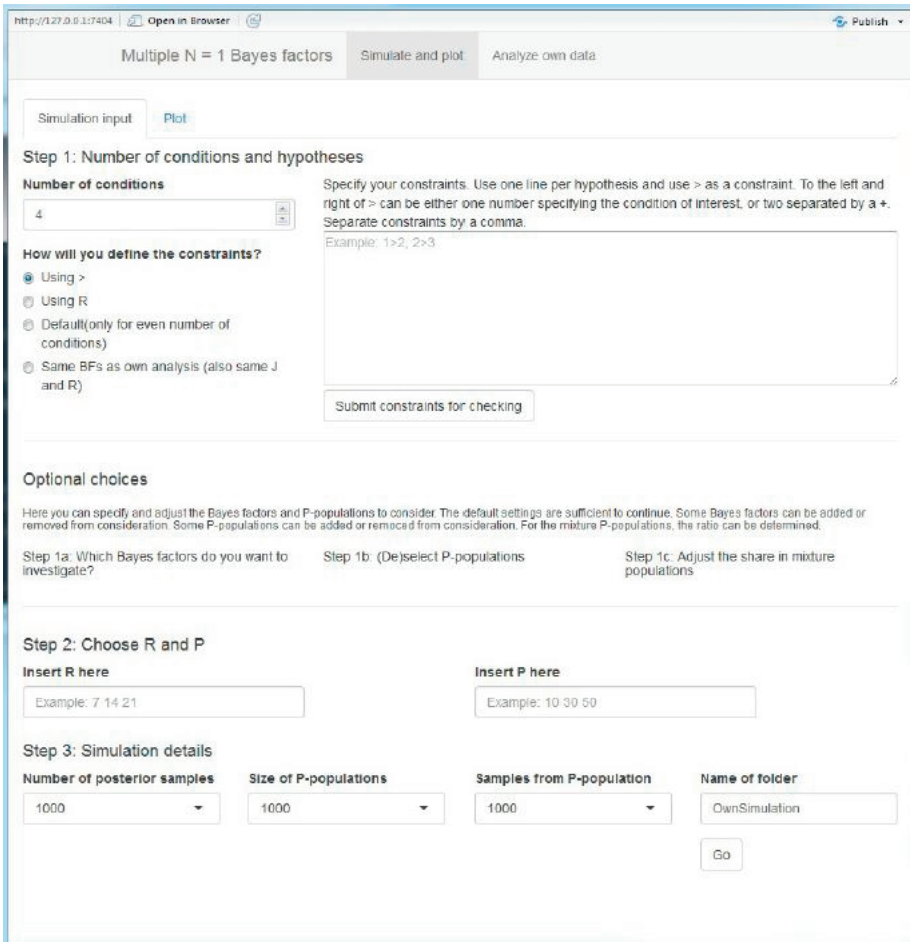


Figure 6.1. Hypothesis specification option 1: Using >

Step 3: Computation details

The number of iterations required for the computation of the Bayes factor. By default, this value is 10,000. Decreasing this value will increase the speed of computation, but particularly for larger number of conditions (say 8), decrease the precision of the Bayes factor computation. You can enter and adjust the computation seed, for reproducibility of your results.

By then pressing the button `Execute analysis' (appears if the constraints are filled in and checked), the computation will start. A pop-up will appear in the bottom right corner to indicate that the computation is busy, and a notificationAnalysis finished' will appear under the button when ready. Then, you can access the other two tabs to view the results

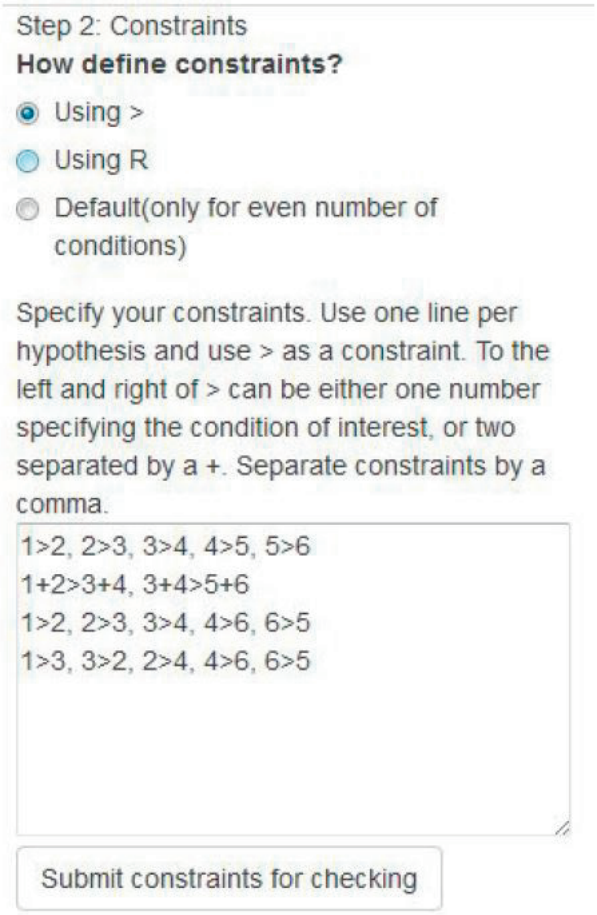


Figure 6.2. Hyptothesis specification option 2: Using *R*

Chapter 7

Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence

by I.J.M. van der Ham, F. Klaassen, K. van Schie, and A. Cuperus[1]

7.1 Introduction

The quality and number of applications of virtual reality (VR) environments are rapidly increasing. VR allows for a controllable approximation of the real, physical world that can be used in a wide range of situations (e.g., for entertainment or medical purposes). Yet, there appear to be limitations to the extent to which the physical world can be imitated. For instance, distance has been found to be underestimated in VR environments (e.g. Knapp & Loomis, 2004; Finnegan, 2016; Stefanucci, Creem-Regehr, Thompson, Lessard, & Geuss, 2015) and the accuracy by which spatial information is perceived can easily be manipulated in VR (e.g. Linkenauger, Bülthoff, & Mohler, 2015; Cuperus & Ham, 2016; Cuperus et al., 2018) Such effects could have substantial impact on experimental and practical implementations of VR, as they may interfere with perceptual processes relevant to the task at hand. Underestimation in VR environments may also extend to the temporal domain, as essential cues supporting time estimation ('zeitgebers') such as the position of the sun are lacking or can easily be manipulated (Schatzschneider, Bruder, & Steinicke, 2016).

[1]Published as Van der Ham, J. M. E., Klaassen, F., van Schie, K., & Cuperus, A. (2019). Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence. *Computers in Human Behavior, 94*, pp.77-81.

Author contributions: IH and AC designed the study and collected the data. IH, FK and AC developed the hypotheses. FK and KvS discussed the statistics. FK analyzed the data and wrote to the methods and results section. IH wrote the final product, FK, KvS and AC provided written feedback.

Several therapeutic applications of VR support a time compression effect; for instance, breast cancer patients underestimated elapsed time after VR-mediated chemotherapy, whereas they overestimated it after music-mediated chemotherapy (Chirico et al., 2016). VR can also be used as a distraction method during medical procedures, in order to relieve pain (Indovina et al., 2018). Thus, VR may be used during stressful procedures like chemotherapy to produce an elapsed time compression effect. It then serves mainly as a distracting circumstance, as it is thought to reduce the overall impact of the medical procedure by making it seem to last shorter. However, the extent of this effect have been found to depend on the type of cancer patient exposed to a VR element in their treatment. Breast cancer patients were more likely to experience altered time perception, whereas lung cancer patients were less likely. The cause of such individual variation remains unclear (Schneider, Kisby, & Flint, 2010). Furthermore, other more exploratory findings suggest a deviation of time perception in the opposite direction, a pilot study making use of a head mounted device found longer perceived elapsed time for the virtual display compared to the real world (Bruder & Steinicke, 2014).

The precise mechanisms underlying such distraction are unclear as of yet. It has been suggested that mainly attentional and affective factors play a role in this process (e.g. Sharar et al., 2016). Such attentional processes could potentially also connect to VR specific time compression effects, analogous to the established spatial underestimation in VR (e.g. Stefanucci et al., 2015). Therefore, the main goal of the current experiment was to determine whether time compression effect exists for VR and if so, which factors of VR presentation cause this effect. A better understanding of the working mechanism of this process could help to optimize future medical interventions based on VR.

So far, studies on time perception in VR are limited and do not reflect on the precise sources of such an effect: is it medium of VR itself that affects time perception, or could it alternatively be caused by the content displayed in VR, as this is often not strictly controlled for in comparisons between real world and VR time perception. Literature concerning temporal processing highlights several factors as key players in distortions in time perception, identical to those mentioned as likely mediators in the process of pain relief by VR (Sharar et al., 2016). Emotion, as expressed by affective valence and arousal level, is of particular importance. In addition, attentional processes are often mentioned in relation to emotion; emotional input draws more attention (Angrilli, Cherubini, Pavese, & Manfredini, 1997; Burle & Casini, 2001; Droit-Volet & Meck, 2007; Matthews & Meck, 2016; Noulhiane, Mella, Samson, Ragot, & Pouthas, 2007). Angrilli et al. (1997) have studied time perception in relation to these factors and found that different patterns of temporal processing are present for different levels of arousal; high arousal stimuli result in shorter time perception and are emotion-driven, whereas low arousal stimuli are linked to longer time perception and appear to be attention-driven.

So, the few VR studies on this matter suggest VR is linked to time compression and would predict that time is perceived to go faster in VR compared to the physical world. As VR has been found to elicit emotional responses (e.g. Felnhoger, 2015), one viable explanation is that VR itself is the cause of distortions in temporal perception. Alternatively, it may be the content of VR presentation that results in the elapsed time compression effect, as this may well differ in level of arousal and emotional valence. Literature suggests that in this case, high arousal stimuli are perceived to go faster than low arousal stimuli (e.g., Angrilli et al., 1997). Therefore, we conducted an experiment comparing time estimation of

videos presented in VR to those presented in the physical world, in a highly similar visual environment. The videos varied in their emotional content, and participants' individual ratings of valence and arousal were included in the analyses.

A better understanding of time perception in VR will not only help understand how humans process virtual environments, but may also clarify how VR can best be used in medical settings such as chemotherapy or other painful procedures. Is it really VR itself that functions as a 'time compressor' or is it the content used, and could these also be presented through a means of presentation other than VR?

7.2 Methods

7.2.1 Participants

Twenty-nine participants took part in the study (15 male, 14 female, mean age = 24.8, SD = 3.13). Exclusion criteria were a self-reported history of psychiatric or neurological disorders, proneness to motion sickness, and visual impairments. The study was approved by the Leiden University Ethical Committee of the Institute of Psychology (CEP16-0309/124).

(a)

(b)

Figure 7.1. Experimental set up in (a) the RL cinema and (b) the virtual rendition of the RL cinema.

7.2.2 Setting and materials

Participants viewed movie clips in a VR setting and in real life (RL). The RL situation for this experiment was a movie theatre (Cinemec in Utrecht, the Netherlands), with a 5 by 9 meter digital cinema projector (DP2K-19B; Barco; Kortrijk, Belgium). Participants were seated in an empty theatre, in a central position to the screen. The images shown in the VR setting accurately resembled this setting; when participants wore the VR headset (Samsung Galaxy S6 + Gear VR; Samsung Electronics; Daegu, South-Korea), they saw the movie screen from the same position, with highly similar colour scheme and lighting (see Figure 1).

In both conditions, participants viewed a series of short movie clips. Two sets of movie clips were created, each with a total duration of 18 minutes, containing 10 different clips of

varying lengths (range: 7-90 seconds). The content of these movie clips was based on the international affective picture system (IAPS; Lang (1997)). Appropriate movie equivalents of the pictures in this system were selected by two of the experimenters, to reach a stimulus set with substantial differences in levels of arousal and affective valence (e.g., crawling spider, starving lion, people fighting, coconut shells).

7.2.3 Task design and procedure

Participants signed the informed consent form and proceeded with filling out a basic questionnaire concerning demographic information. Then, they were instructed to put away any watches or phones or devices with a clock before starting the experiment. Participants were then shown a set of movie clips in either the real life movie theatre setting or the VR environment. After each clip a blank screen appeared for 60 seconds, during which they were asked to estimate the duration of the clip in seconds. For each movie clip, the difference between the estimated time (ET) and actual time (AT) was computed, and divided by the actual time to compensate for the difference in actual time of the clips. This provides the relative difference (RD) in time estimation: RD = (ET – AT) / AT, where RD = 0 indicates the estimated time was equal to the actual time, positive scores indicate the proportional overestimation of actual time (i.e. time compression), and negative scores indicate the proportional underestimation of actual time (i.e. time expansion). Furthermore, participants rated level of arousal and affective valence they experienced while viewing the clip on a Likert scale ranging from 1 *calm/very negative* to 9 *aroused/very positive* (see Angrilli et al., 1997).

Each participant viewed both sets of movie clips; one in the RL setting and one in VR. Participants were evenly distributed across the four experimental conditions, with the two types of environment and two sets of movie clips combined in pseudorandomized order.

7.2.4 Statistical analyses

The main interest of this study is the effect of condition (VR vs RL), arousal, and valence of movie clips on the relative difference in time perception. This can be analysed by means of a regression analysis. However, the data contain a dependency within participants: the measurements for different movies are nested within the participants (i.e., each participant responds to multiple movies). Therefore, we analysed the data using a multilevel model that can account for this dependency. The model was specified as follows:

$$\begin{aligned}\text{Relative Difference in Time Perception}_{im} = & b_{00} + b_{0c}condition_{im} + b_{0a}arousal_{im} + \\ & b_{0v}valence_{im} + u_{i0} + e_{im},\end{aligned}$$

In this model the *Relative Difference in Time Perception* for person i and movie m is explained by a grand intercept (b_{00}), with individual variation (u_{i0}, random intercept), the condition in which person i watched movie clip m ($condition_{im}$, 0 = *RL*, 1 = *VR*), the subjective level of arousal of the movie m ($arousal_{im}$) and subjective affective valence of the movie ($valence_{im}$), and the residual error (e_{im}). Note that the main difference

with a normal regression is that in the current model a random intercept ui0 is included. This parameter accounts for individual differences in how people estimate time duration: One person might generally overestimate duration, while another person might generally underestimate time duration, but the effect of condition, arousal and valence can still affect their personal baseline score similarly. Finally, rather than estimating this individual effect for every person, a multilevel model assumes that these individual deviances from the grand mean/intercept are normally distributed, with a mean of 0, and a variance τ_u^2. If this variance is 0, there is no individual variation.

Using the model above, we tested three informative, competing hypotheses:

$$H_1 : b_{0c} = 0, b_{0a} > 0, b_{0v} > 0$$
$$H_{1c} : \text{not } H_1$$
$$H_2 : b_{0c} > 0, b_{0a} > 0, b_{0v} > 0$$

H_1 expresses that there is no effect of condition on the relative time estimation (i.e., time estimation for VR and RL are similar), and that both arousal and valence have a positive effect on relative time estimation (higher scores on valence/arousal correspond to a stronger overestimation of movie clip duration). H_{1c} is the complement of H_1, which means that it encompasses all other possible combinations of the parameters in H_1. Finally, H_2 specifies the same effects of arousal and valence, and additionally that the VR condition results in larger relative time perception scores. We are interested in comparing H1 with H1c to learn whether this model is better than its complement and comparing H_1 with H_2 to test the effect of condition. These hypotheses are not in the traditional format of null and alternative hypotheses. They are more specific and can be considered 'informative hypotheses' (Hoijtink, 2012). These hypotheses cannot be evaluated with frequentist analyses, and therefore a Bayesian model was adopted. This makes for two substantial differences compared to more standard analyses. First, a prior distribution has to be specified for all parameters. Second, the Bayesian evaluation of hypotheses does not result in p-values, but in two Bayes factors quantifying the relative evidence for H_1 versus H_{1c} and for H_1 versus H_2. Both these elements will be discussed in more detail in the results section.

7.3 Results

The hypotheses of interest cannot be compared to one another using frequentist statistical analyses. Bayesian methods allow for the comparison of the specified hypotheses. We used the Bayesian software Bain (Gu et al., 2017; Hoijtink, Gu, & Mulder, n.d.) that is designed to evaluate hypotheses that may consist of inequalities (larger, smaller than) and equalities between parameters. Bayesian analyses require the specification of a prior distribution for the parameters. The software Bain computes a minimally informative prior distribution using a minimal training sample of the data (Hoijtink et al., n.d.). This minimal training sample is based on the estimates and covariance matrix of the relevant parameters. To obtain these estimates the multilevel model was run using JAGS version 4.3.0 (Plummer, 2003) in R version 3.4.2 (R Core Team, 2013) with vague priors (see Appendix 10.3.1 for the full JAGS code, including the prior distributions used).

Table 7.1
Parameter estimates

Parameter	HPD estimate	95% CI	Standard error	Std. coefficient
b_{00}	-0.241	[-0.440 : -0.043]	.100	-.221
b_{0a}	0.009	[-0.009 : 0.028]	.010	0.019
b_{0c}	0.014	[-0.056 : 0.084]	.036	0.028
b_{0v}	0.010	[-0.011 : 0.030]	.010	0.019
τ_e	5.164	[4.570 : 5.793]	.313	1.368
τ_u	14.038	[7.151 : 23.872]	4.311	3.756

Note. Highest posterior density parameters estimates obtained from the Bayesian analysis, with a 95% Credible Interval, standard error and standardized parameter value. b_{00} denotes the intercept, b_{0a}, b_{0c} and b_{0v} the regression coefficient for arousal, condition and valence, respectively, τ_e denotes the residual variance and τ_u the individual intercept variance.

Table 7.1 presents the Highest Posterior Density (HPD) estimates of the parameters in the model (Bayesian equivalent of parameter estimates) along with the 95% Credible Interval (Bayesian equivalent of confidence interval) and the standardized regression coefficients. This table shows that there is reason to believe that the intercept is indeed random; the variance of the random effect (u_{i0}) is larger than 0, indicating that individuals differ in their average time perception. Furthermore, it is evident that condition is the strongest predictor for time perception through comparing the standardized regression coefficients.

In addition to the hypotheses, estimates and estimated covariance matrix, Bain requires the sample size. The sample size determines the fraction of information taken from the data to compute the prior distribution (Hoijtink et al., n.d.). The available data consist of 20 repeated measures for each of the 29 individuals, resulting in a total of 580 data points. These data points do not all contribute unique information because they are nested in the 29 individuals. Computing the prior distribution using a sample size of 580 would unfairly assume we had 580 independent pieces of information. The sample size should be somewhat smaller than 580. If no variation existed among the measurements in each participant, the effective sample size would be 29. Simulations researching power in multilevel models tell us that observed power is a function of both the number of clusters and the number of measurements (e.g., Maas & Hox, 2005; Scherbaum & Ferreter, 2008). The effective sample size is between the number of clusters (29 individuals) and the number of measurements (580).

We executed the analysis for different choices of sample $N_{effective} = 29, 180, 380, 580$. The minimum considered sample size of 29 reflects the sample size if no variation existed in within-person measurements. This can be considered a 'worst case scenario': the computed prior contains very little information and estimation because fairly unstable. The maximum considered sample size reflects the sample size if there is no between-person variation. This choice would overfit the estimation, because any between-person variation is not accounted for. The sample sizes of 180 and 380 are the sample sizes we consider to reasonable reflect the within-between person variance balance. By considering this range of sample sizes for the computation of the prior distribution, we can compare the results and evaluate the impact of the dependency on the results.

Table 7.2
Bayes factors

Effective sample size	29	180	380	580
H_1 vs. H_{1c}	8.53	21.25	30.87	38.14
H_1 vs. H_2	2.23	5.54	8.05	9.95

Note. Bayes factors expressing the relative evidence in the data for H_1 versus H_{1c} (top row) or H_2 (bottom row) for effective sample sizes 29, 180, 380 and 580. Bayes factors for the unstandardized analysis are presented here. Bayes factors are similar for the standardized analysis.

Table 7.2 shows the Bayes factors that describe the evidence in the data for H_1 relative to H_{1c} and H_2. Both BF_{1c} and BF_{12} increase as the effective sample size increases. The direction and strength of the evidence is rather stable for $N_{effective} = 180, 380, 580$. Both BF_{1c} and BF_{12} are considerably weaker only for $N_{effective} = 30$. The sensitivity analysis shows that for the more reasonable effective sample sizes, strengths of evidence are similar.

The hypothesis that there is no effect of condition, in combination with an effect for arousal and valence (H_1), is supported over its complement (in the first row in Table 7.2 the Bayes factor is always larger than 1, indicating that H_1 is $8.53/21.25/30.87/38.14$ times more supported than H_{1c}) , and is preferred over H_2 where there is an effect of condition (presented in the second row in Table 2).

Note that other than the within-participant dependency, there is an additional dependency within the clips viewed (i.e., for half of the participantsthe first set of movie clips was presented in the VR condition, and the second set in the RL condition, and vice versa for the the other half of the participants). This might create noise in the analysis if a particular clip is structurally rated higher in the VR condition than in the RL condition or vice versa. The fragments in each set of clips were selected to be similar, so the expected effect of this dependency should be small or negligible. To check whether there was dependency within movies, the hypotheses were evaluated in a more elaborate model that accounts for the within-movie dependency in addition to the within-person dependency. For every movie, a random intercept is included in the model. This model resulted in very similar results (see Appendix 10.3.2 for the more elaborate model and the results).

7.4 Discussion

The use of VR is rapidly increasing in a range of applications, including clinical treatment protocols. One characteristic of VR use in clinical context is that it is claimed to result in compressed time perception, yet evidence is limited and the potential source of such temporal compression is unclear. Analogous to compression found in the spatial domain, the virtual display itself could be the cause. Alternatively, the affective nature of the content displayed in VR may cause temporal compression. In this study we first addressed the question whether time is perceived to pass by faster in VR. Next, we examined if such an effect wasrelated to the medium of VR itself, or the content of the materials used, in terms of emotional valence and arousal. Given the characteristics of the dataset, a Bayesian

approach was used in which 3 hypotheses were tested and consequently compared based on the evidence. The hypothesis with the strongest relative evidence was that both arousal and valence positively contribute to the observed time compression effect, regardless of the viewing condition. Thus, there is no evidence for a difference in temporal processing between VR and RL. So, when filtering out the impact of the content of stimuli, the medium of VR itself does not affect time perception in our experiment. Furthermore, this finding suggests that the time compression effect that takes place is most likely the result of the emotional content of the materials displayed. This finding is in line with Angrilli et al. (1997), as higher arousal is linked to shorter time perception. Moreover, this would also mean this process is mainly emotion-driven, not attention-driven, given Angrilli's (1997) description of the characteristics of higher arousal. This finding is analogous to a potential explanation for how VR may cause pain relief during medical interventions, which has been suggested to rely on affective factors (Sharar et al., 2016).

Reports on reduced time perception within clinical contexts, where unpleasant clinical procedures are performed when VR is employed do not necessarily conflict with these findings. As those comparisons typically use different visual materials in the VR condition, the emotional content participants are exposed to also differs between the VR and RL conditions. The current experiment's set up uniquely allowed for a direct comparison, as it made use of a VR environment highly similar to the RL environment, with identical video materials.

It should be noted that the analyses do not allow for a distinction between negative and positive emotional valence, as valence was represented as a continuous scale instead of a dichotomy. Other limitations of the current study concern the demographics of the participants; possibly gender has and effect (Hancock & Rausch, 2010) and age range in particular may be different in clinical populations in which such VR interventions are used and could therefore be considered in future research.

The current study taps into a relatively new area: how time is perceived when engaging in virtual environments. This has implications for both experimental and clinical context. The use of VR is increasingly popular in cognitive experiments and is often considered a reliable source of information concerning human behavior in the real world. Yet, the current data suggests that some caution is warranted. Even though the medium itself does not affect how time is perceived, the emotions evoked by the stimuli at hand may cause a difference. This could affect measures of time-related cognitive abilities, such as episodic memory. In clinical context, this shows that it may be possible to achieve the desired time compression effects through other means than VR, as the main cause appears to be the affective content rather than the medium itself. Future research should be directed at isolating the contributions of negative and positive valence, and other formats of stimulus display.

7.5 Conclusion

The current findings shed light on how humans temporally process virtual environments: this process is highly similar to that in RL. The emotional content of the materials used affects temporal processing, regardless of condition. This may contribute to the implementation

of VR in therapeutic settings, as VR itself may not be necessary to achieve the desired time compression effect during medical procedures. To this aim, future research could be directed at separating the roles of negative and positive emotional valence.

Chapter 8

Using Bayesian methods to test mediators of intervention outcomes in Single case experimental designs (SCEDs)

by M. Miočević, F. Klaassen, G. Geuke, M. Moeyaert, and M. Maric.[1]

8.1 Introduction

8.1.1 Single-case experimental designs (SCEDs)

Single-case experimental designs (SCEDs) methodology is a rigorous scientific research approach that can be used to evaluate the effectiveness of an intervention (Horner et al., 2005; Kazdin, 2011). SCEDs have shown to be a prime alternative for large-group studies either as an initial study leading to specific hypothesis to be tested in a group study, or as a stand-alone research study. This second option is especially important in heterogeneous populations, or populations with rare incidence rates which may not be uncommon in communication disorders research. Since SCEDs can also easily be incorporated in clinical practice, they have the potential to enhance evidence-based practice and stimulate collaboration between research and practice, unifying research questions that

[1]Manuscript submitted to Evidence-Based Communication Assessment and Intervention

Author contributions: MMa identified the need for methods to test mediation in SCEDs, and wrote parts of the introduction and discussion. MMi and FK developed the methods in the paper and wrote the methods and results section, and the annotated syntax in the appendix. MMo wrote the section on SCEDs, identified a suitable data set, and provided feedback during the development of the methods. GG wrote the section on data, plotted the data, and contributed parts of the introduction.

emerge from clinical practice on one hand, and, on the other hand, research methodology to test these questions on a single-client level.

The ultimate goal of SCEDs research methodology is to evaluate whether there is a functional relationship between the intervention and change in the outcome measure of interest. For this purpose, a case is measured repeatedly over time during a baseline condition that is 'interrupted' by an intervention (also referred to as "treatment" in the remainder of the paper). By using SCEDs methodology, a case serves as its own control, detailed information related to changes across time can be obtained, and case-specific intervention effects can be estimated. Because of these advantages, the design is becoming increasingly popular over time and it has been the method of choice for over a thousand studies to date (Wiessenekker, 2019). SCEDs are used across a variety of different research fields ranging from rehabilitation and clinical psychology to special education, and are known under a several different names such as interruinterrupted time series, single-subject experimental design, intrasubject designs, $N = 1$ designs, etc. (Smith, 2012).

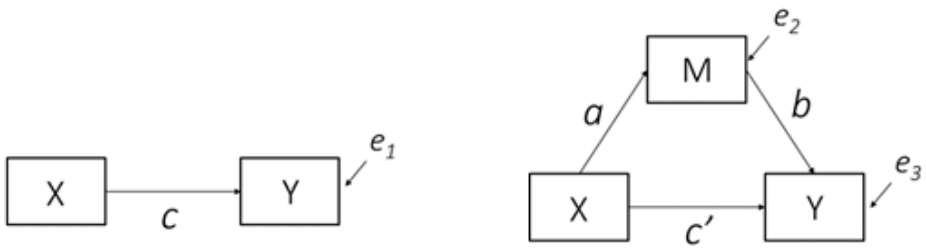
Together with the increasing interest in using SCEDs to establish an evidence base for the effectiveness of treatments, there is a need for methods to quantify the size of the intervention effect. During the last decade there have been efforts to develop and empirically validate indices and effect sizes to report the strength and statistical significance of effects. However, there is no best index (What Works Clearinghouse, 2017) and some indices might be better in some conditions compared to others (Manolov & Moeyaert, 2016; Vannest, Peltier, & Haas, 2018). Non-parametric nonoverlap indices quantify the degree of non-overlap between the baseline and the treatment data clouds, such as Non-overlap of All Pairs (NAP; Parker & Vannest, 2009), Tau-U (Parker, Vannest, Davis, & Sauber, 2011), Tau-C (Tarlow, 2016), Improvement Rate Difference (IRD; Parker et al., 2009). and the Percent of Data Exceeding the Phase A Median Trend (PEM-T; Wolery, Busick, Reichow, & Barton, 2008) just to name a few. Parametric approaches on the other hand allow for a quantification of the size of a treatment effect together with an estimate of the standard error. Some popular parametric approaches are regression-based effect sizes (i.e., Center, Skiba, & Casey, 1985; van den Noortgate & Onghena, 2003a, 2003b), multilevel modeling (Shadish, Rindskopf, & Hedges, 2008), hierarchical linear modeling (Parker et al., 2009), standardized mean differences (e.g., Cohen's d, Hedge's g; Shadish, Hedges, & Pustejovsky, 2014) and the between-case standardized difference (Hedges, Pustejovsky, & Shadish, 2012, 2013). All of these approaches can be used to test the effectiveness of a therapy, i.e., provide an answer to a 'yes/no' question: 'Does the treatment work for this individual client?'

8.1.2 Moving beyond the 'yes/no' question: Mediation analysis

Nowadays, personalized medicine is becoming more popular, and we are aware that interventions that work for one person may not work for another person, or may work for multiple participants, but due to different causal mechanisms. When studying effects on a group level, scientists implicitly assume that interventions work the same for all group members, and neglect the unique reasons why certain interventions work (or do not work) for clients. Without examining effects at the individual level, we cannot evaluate the causal mechanism through which a treatment works (or does not work) for a given

person. Generalizing causal relationships from the group-level to the individual level is not recommended (Cattell, 1952).

Mediation analysis is used to evaluate intermediate variables (mediators; M) that transmit the effect of an independent variable (X) on a dependent variable (Y) (MacKinnon, 2008). It provides an answer to a question: "How does the treatment work, through which mechanisms?" For example, Maric and colleagues (Maric, Heyne, MacKinnon, Widenfelt, & Westenberg, 2012) found that self-efficacy mediated the relationship between cognitive-behavioral therapy (CBT) and school-related fear in adolescents. Thus, the theory tested by mediation analysis in clinical settings is that a certain intervention will produce changes in the mediator and that these changes will, in turn, affect intervention outcomes (MacKinnon, 2008). So far, these intervention theories have, unfortunately, only been tested in large-group studies. In the remainder of this section we describe a single mediation model (see Figure 8.1) and the most frequent data-analytic approaches to testing for mediation.



(a) Total effect of the independent variable on the outcome

(b) Single mediator model

Figure 8.1. Mediation models. The intercepts are included in the two models, but not in the figure.

The effects of interest in the single mediator model (Figure 8.1) can be computed using three equations:

$$Y = i_1 + cX + e_1 \tag{8.1}$$

$$M = i_2 + aX + e_2 \tag{8.2}$$

$$Y = i_3 + c'X + bM + e_3 \tag{8.3}$$

where X is the independent variable, M is the mediator, and Y is the dependent variable. Intercepts are i_1, i_2, and i_3, c is the total effect of the independent variable on the dependent variable, a is the coefficient relating the independent variable to the mediator, b is the coefficient relating the mediator to the dependent variable in the model containing the independent variable, c' is the coefficient relating the independent variable to the dependent variable (also called the direct effect), and e_1, e_2, and e_3 are error terms assumed to follow a normal distribution with a mean of 0 and variances of σ^2_{e1}, σ^2_{e2} and σ^2_{e3} respectively.

One of the first approaches to testing for mediation was described in papers by Judd & Kenny (1981) and Baron & Kenny (1986), and it consists of four steps: (1) establishing that the independent variable affects the dependent variable (i.e., significant coefficient c

in Equation 8.1); (2) establishing that the independent variable affects the mediator (i.e., significant coefficient a in Equation 8.2); (3) establishing that the effect of the mediator on the outcome, controlling for the independent variable, is nonzero (i.e., significant coefficient b in Equation 8.3); (4) establishing that the effect of the independent variable on the dependent variable is weaker when we control for the effect of the mediator than when we do not control for the effect of the mediator (i.e., coefficient c' in Equation 8.3 should be smaller than coefficient c in Equation 8.1). This approach falls under the category of causal steps approaches to mediation analysis, and one of the less stringent and more powerful causal steps methods is called the joint significance test, which only requires Steps 2 and 3. However, none of the causal steps approaches provide a numerical estimate of the value of the indirect (mediated) effect, and they have less power to detect the mediated effect relative to methods that compute and test the significance of the mediated effect directly (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002).

The mediated (indirect) effect is most often computed as the product of coefficients ab, and in linear models with no missing values, we obtain the same value of the mediated effect if we compute it as the difference of coefficients $c - c'$ (MacKinnon, Warsi, & Dwyer, 1995). Modern approaches to mediation analysis test the significance of the mediated effect by computing confidence intervals for the mediated effect and evaluate whether 0 is in the interval. Modern methods that have the most power either model the distribution of the mediated effect appropriately (i.e., using the distribution of the product of two normal variates; Craig, 1936; Lomnicki, 1967; MacKinnon et al., 2002; MacKinnon, Lockwood, & Williams, 2004) or do not make any assumptions about the distribution of the mediated effect (e.g., bootstrap and Bayesian methods; MacKinnon et al., 2004; Yuan & MacKinnon, 2009).

8.1.3 Bayesian mediation analysis

The mediated effect can be computed and evaluated in the frequentist (classical) framework using methods such as ordinary least squares regression or structural equation models fit using Maximum Likelihood estimation. It is also possible, and sometimes more advantageous, to do mediation analysis in the Bayesian framework (Miočević, MacKinnon, & Levy, 2017; Yuan & MacKinnon, 2009). In the Bayesian framework, the analysis starts by specifying prior distributions for all freely estimated parameters in the model. In the case of the single mediator model, the parameters that are assigned priors are those from Equations 8.2 and 8.3: the intercepts i_2 and i_3, regression paths a, b, and c', and residual variances σ^2_{e2} and σ^2_{e3}. The next step of a Bayesian analysis requires updating the prior distributions with the observed data using Bayes' theorem, in order to obtain the posterior distribution of the model parameters: $p(\theta \mid data) \propto p(data \mid \theta)\, p(\theta)$, where $p(\theta \mid data)$ denotes the posterior distribution of the parameters, $p(data \mid \theta)$ denotes the likelihood function based on the observed data, and $p(\theta)$ denotes the prior distribution for the set of freely estimated parameters. The inferences about the parameters of interest are based on the posterior distributions that can be summarized to obtain a point summary (e.g., mean or median) or an interval summary. The distribution of the mediated effect is approximated using values from the posterior distributions for coefficients a and b. These distributions can be obtained using Markov Chain Monte Carlo (MCMC), implemented in various software (for a tutorial on using MCMC, see Sinharay, 2004). The MCMC draws can

be used to approximate the posteriors, but also for hypothesis testing. Bayesian statistics have a unique take on hypothesis testing, and allow for quantifying relative evidence for different hypotheses using a Bayes factor (Kass & Raftery, 1995). Bayesian hypothesis testing is very flexible in terms of hypotheses that can be compared. Expectations about the directions of the effect (e.g., the sign of a regression coefficient) can be formulated as so-called *informative* hypothesis (Klugkist et al., 2005). This allows both for the inclusion of expectations about the directions of effect in the hypothesis, and for testing multiple effects simultaneously. Additionally, more than two hypotheses can be compared at the same time, thus allowing for the selection of the best hypothesis out of the entire set. For the sake of space, we cannot provide a more extensive description of Bayesian methods for mediation analysis and informative hypothesis testing, and we refer the interested reader to chapters by Miočević (2019), the paper by Yuan & MacKinnon (2009), the book by Hoijtink (2012) and the paper by Béland, Klugkist, Raîche, & Magis (2012).

The above methods are frequently used for group-level (i.e., studies with $N > 1$) mediation analyses. There has been only one proposed method for mediation analysis in context of SCEDs (Gaynor & Harris, 2008). However, the proposed method does not yield a numerical estimate of the mediated effect, nor does it allow the researcher to quantify the support of the mediation hypothesis from the data. Knowledge about individual participants' mediators of treatment outcomes could inform treatment-decision making and lead to a more evidence-based practice (Maric, Prins, & Ollendick, 2015). Furthermore, knowing the mediator(s) that transmit the effect of an intervention on the outcome(s) of interest can help in tailoring the treatment to each patient.

8.1.4 This study: SCEDs meeting mediation analysis

In this paper, we describe two methods for evaluating whether there is a mediated effect: a method that can compute the value of the mediated effect using repeated measures of a hypothesized mediator and an outcome of interest collected from a single participant, and a method that tests whether this mediated effect is different from 0. The methods developed and described in this paper will use Bayesian estimation for the parameters in the mediation model, and this is the first paper (to our knowledge) that includes both parameter estimation and informative hypothesis testing for mediation models. We will focus on the regression-based effect size originally introduced by Center et al. (1985) because of its flexibility. In order to estimate the regression-based effect size, a piecewise regression can be run which results in the estimate of the outcome score at the start of the SCED, the time trend during the baseline, the immediate intervention effect (i.e., change in outcome score at the start of the intervention phase) and the difference in time trend between the baseline phase and the intervention phase. This results in two regression-based effect sizes of interest, namely an immediate intervention effect and an intervention effect on the time trend.

The following sections describe the data for the empirical example and how Bayesian piecewise regression analysis can be used to test for mediation in a SCED.

8.2 Empirical example

8.2.1 Data

The dataset for the empirical example comes from a study of the effectiveness of wearing the Playskin Lift™exoskeletal garment on object exploration and cognitive outcomes in infants that were born preterm and/or had brain injuries (Babik et al., 2019). The exoskeletal garment was designed to assist antigravitational movement of the infant, which was hypothesized to aid object grasping and exploration. For a more detailed and comprehensive description of the dataset and measurement procedure of this study, the reader is referred to the article by Babik and colleagues (2019).

The dataset is a multiple baseline $A_1B_1A_2$-design, which means that it consists of three phases: the first phase is a baseline phase (A_1), which was designed to assess the baseline level of the infant's scores on various variables of object exploration and reaching. The exoskeletal garment was only worn during a subset of assessments in this phase. The amount of measurement occasions in this baseline phase was alternated across participants, ranging from 3 to 5 occasions. The second phase (B_1) is the treatment phase, in which parents were asked to perform a structured set of daily exercises of 40 minutes with the infants using the exoskeletal garment. The third phase (A_2) was a follow-up phase, which was designed to assess whether there were remaining effects of using the exoskeletal garment after the treatment was stopped, and was similar to the baseline phase. As mentioned before, because the effect of the intervention on the outcome score is replicated across multiple participants, the SCED study is more externally valid (i.e., more generalized conclusions about the intervention effectiveness can be obtained).

At each measurement occasion, six types of assessments were conducted. Each assessment consisted of a toy presentation to the infant, after which the reaction of the infant was measured in a structured manner. This assessment was conducted in 2x3 conditions, both with the exoskeletal garment off and on, and with the toy presented at hip, chest, or eye level. All assessments were recorded on video. For each of these assessments, several variables were recorded, such as grasping ability and the percentage of time the infant looked at the toy.

For the purposes of the current example, a subset of the variables of one participant will be used to illustrate the suggested analysis methods. The mediation hypothesis was that daily exercise with the exoskeletal garment (X; treatment) leads to better grasping ability (M; mediator), which leads the infant to be more interested in toys and more time spent looking at the toy (Y; outcome). Grasping was measured as the percentage of the total assessment time in which the infant had any type of contact with the toy, i.e. the sum of bimanual and unimanual contact. Looking was measured as the percentage of the total assessment time in which the infant directed their eyes at the toy. Data for the empirical example are plotted in Figure 8.2. One condition of measurements was selected for the illustrative analysis here: with the exoskeletal garment off and the toy presented at the chest level. Arguably, the aim of the treatment in the study by Babik and colleagues (in press) was to improve the independent grasping abilities of the infants, i.e. without wearing the exoskeletal garment. Note that, for a more complete analysis of this data, the proposed analysis can be repeated for all six conditions and that the methods we illustrate use only the baseline phase (i.e., A_1)

and the intervention phase (B_1), but could be extended to include additional phases (e.g., A_2, which presents the maintenance phase in the present data set). Also note that using data of only one participant of a multiple baseline study does not allow the analysis to make generalizations (i.e., external validity) about the intervention and the mediation effect.

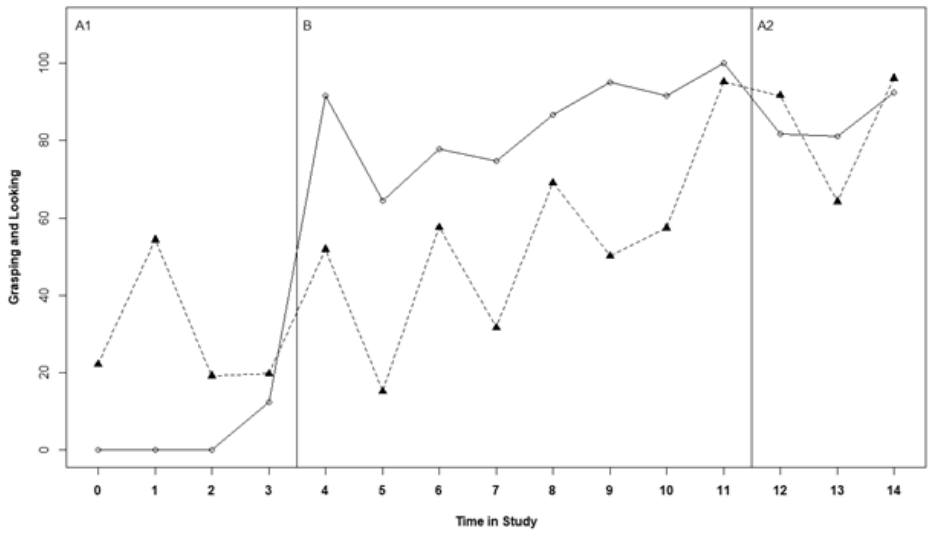


Figure 8.2. Graphical display of the scores of Grasping (dashed lines and triangles) and Looking (solid lines and points) of participant 201 of the study by Babik et al. (in press). Phases are denoted in the upper left corner of each phase.

For readers interested in using the example code provided in Appendix A, it is important to organize the data in a specific format for the code to work. The data set needs to contain the following variables: (1) Time, which denoted the measurement occasion and in the current analysis ranges from 1-12; (2) Phase, which denotes whether a given observation belongs to the baseline phase (Phase = 0) or the treatment phase (Phase = 1); (3) Time1, which is equal to the value of Time - 1 (and ranges from 0 to 11 in the present data set); (4) Time2, which is equal to the variable Time1 minus the number of measurement occasions in the baseline phase and for which a value of 0 denotes the start of the treatment phase (in this data set, Time2 ranges from -4 to 7); (5) phase_time2, which denotes the time spent in the treatment phase, and has a value of 0 during the baseline phase and at the first occasion in the treatment phase, and values of 1, 2, 3, etc. for subsequent observations in the treatment phase; (6) ScoreM, which are scores on the mediator on occasions 1-12; (7) ScoreY, which denotes the score on the outcome at a given measurement occasion (in the present data set, there are 12 values of ScoreY); (8) Tmed, which represents scores on the mediator with a missing value in the first row and scores on occasions 1-12 as values in the subsequent rows; and (9) Tout, which are scores on the outcome variable with a missing value in the first row, and scores on occasions 1-12 as values in the subsequent rows. The current formatting of the data set will yield a data set with the number of rows equal to the number of observations plus one; also, variables Time, Phase, Time1, Time2, phase_time2,

ScoreM, and ScoreY will be missing a value in the last row, while variables Tmed and Tout will be missing a value in the first row. This data format is necessary for executing the analyses for the proposed methods.

8.2.2 Methods

The majority of data analytic methods for SCEDs were developed with the goal of evaluating the effect of a change in phase on a single variable. In the single mediator model for SCEDs, both the hypothetical mediator and outcome are measured repeatedly over at least two phases (i.e., baseline phase and intervention phase). Given that our goal is to compute the numerical value of the indirect effect, we automatically excluded methods that quantify percentage of nonoverlapping data (Scruggs, Mastropieri, & Casto, 1987). We opted for piecewise regression analysis instead, as it allows for quantifying the change in the mediator due to the change in phase (a-path in Figure 8.1) and change in outcome due to the change in the mediator (b-path in Figure 8.1) controlling for the effect of phase. For the purposes of the current analyses, the equations for piecewise regression analyses of the mediator and outcome are as follows:

$$M = b_{0M} + b_{1M}time1 + b_{2M}phase + b_{3M}phase\_time2 + b_{4M}M_{t-1} + e_M \tag{8.4}$$

$$Y = b_{0Y} + b_{1Y}time1 + b_{2Y}phase + b_{3Y}phase\_time2 + b_{4Y}M_{t-1} + b_{5Y}Y_{t-1} + e_Y \tag{8.5}$$

Due to the specific coding of the predictors, regression coefficients from the piecewise regression analysis provide estimates of the level of the first time point of phase A (b_{0M} for the mediator and b0Y for the outcome), of the trend in phase A (b_{1M} for the mediator and b1Y for the outcome), of the change in level at the start of phase B (b_{2M} for the mediator and b_{2Y} for the outcome) and of the change in trend between the two phases (b_{3M} for the mediator and b_{3Y} for the outcome; Manolov et al., 2016). The additional terms in the equation for the outcome represent the autoregressive (b_{4M}) and lagged effects of the mediator (b_{4Y}) controlling for the autoregressive effect (b_{5Y}).

There are two reasonable definitions for the effect of the treatment on the mediator (a path in Figure 8.1) in this context: the effect of phase change can either be measured as the change in level (b_{2M}), or as the change in trend between the two phases (b_{3M}). Defining the a path as the change in level between phases allows for computing the indirect effect of the phase change on the outcome through changes in the level of the mediator. Defining the a path as the change in trend between two phases leads to an indirect effect that quantifies the effect of change in phase on the outcome through change in the trend of the mediator. The effect of the mediator on the outcome (b path in Figure 8.1) is represented by the b_{4Y} coefficient from Equation 5 and the direct effect (c' path in Figure8.1) of phase on the outcome controlling for the effect of the mediator is represented either by coefficient b_{2Y} (if the direct effect is defined as a change in level) or using the coefficient b_{3Y} (if the direct effect is defined as the change in trend).

There are two ways to conceptualize the mediated effect in the present example: 1) as the product of coefficients $b_{2M}b_{4Y}$ which represents the change in the value of the outcome due to the change in the level of the mediator following a change in phase, and 2) as the product of coefficients $b_{3M}b_{4Y}$ which represents the change in the value of the outcome due to the

Table 8.1

Ordinary least squares estimates of parameters in Equations 8.4 and 8.5 for Grasping (M) and Looking (Y) and priors for the Bayesian analysis based on these results.

Parameter	Estimate	Standard Error	p-value	Prior
b_{0M} (Intercept)	93.581	20.773	0.004	$\mathcal{N}$(93.581, 1000)
b_{1M} (Time1)	-18.515	8.755	0.079	$\mathcal{N}$(-18.515, 1000)
b_{2M} (Phase)	34.110	20.515	0.147	$\mathcal{N}$(34.110, 1000)
b_{3M} (phase_time2)	28.541	9.081	0.020	$\mathcal{N}$(28.541, 1000)
b_{4M} (Tmed)	-0.797	0.247	0.018	$\mathcal{N}$(-0.797, 1000)
b_{0Y} (Intercept)	-7.934	7.919	0.362	$\mathcal{N}$(-7.934, 1000)
b_{1Y} (Time1)	6.167	3.270	0.118	$\mathcal{N}$(6.167, 1000)
b_{2Y} (Phase)	78.672	9.026	<0.001	$\mathcal{N}$(78.672, 1000)
b_{3Y} (phase_time2)	0.192	3.423	0.957	$\mathcal{N}$(0.192, 1000)
b_{4Y} (Tmed)	-0.009	0.103	0.935	$\mathcal{N}$(-0.009, 1000)
b_{5Y} (Tout)	-0.431	0.099	0.007	$\mathcal{N}$(-0.431, 1000)

Note. The coefficients in the table correspond to the coefficients in Equations 8.4 and 8.5, and the variable names in parentheses correspond to the labels in R output. The symbol N denotes a normal prior distribution where the first parameter represents the mean and the second parameter represents the variance. The analyses were run in rjags so the sample code contains the precision parametrization meaning that the second parameter in the normal priors is the precision and the residual precisions are assigned Gamma (G) priors with both hyperparameters equal to .5.

change in the trend (slope) of the mediator following a change in phase. The procedures for evaluating whether these indirect effects are different from 0 require approximating the distributions of $b_{2M}b_{4Y}$ and $b_{3M}b_{4Y}$, and covariances between b_{2M} and b_{4Y} and between b_{3M} and b_{4Y}, which was more straightforward to obtain in the Bayesian framework. The mediated effect is evaluated using two approaches: parameter estimation and hypothesis testing. Both analyses were performed in R (R Core Team, 2013) using the packages rjags (M. Plummer, 2018) and the software JAGS (Plummer, 2003) for the Bayesian piecewise regression, the R package coda for the computation of intervals for the mediated effects (B. Plummer M., 2018), and the R package bain for hypothesis testing (Gu et al., 2019). The annotated R syntax for the analysis is available in Appendix A. The analysis consisted of the following five steps. Step 1-3 are preparation for Step 4 (parameter estimation) and Step 5 (hypothesis testing):

Step 1. Obtain frequentist estimates of the parameters in Equations 8.4 and 8.5 using the `lm()` function. The estimates and standard errors are shown in Table 8.1.

Step 2. Formulate priors for the parameters in the Bayesian estimation of the parameters in Equations 8.4 and 8.5. These priors have data dependent mean hyperparameters and variance hyperparameters that are diffuse for the scale of the variables (as shown in the last column of Table 8.1). In other words, the priors for each intercept and regression coefficient encode the assumption that the best guess for these parameters is equal to the OLS estimate of that parameter, and the prior variances indicate limited confidence in these best guesses. Data dependent priors are somewhat controversial because they lead to an underestimation of the uncertainty of the parameter estimate/posterior summary (see e.g. Darnieder, 2011).

However, in this situation, fitting the model with normal priors centered at 0 for each intercept and regression coefficient leads to posterior means and medians that are noticeably lower in absolute value relative to the frequentist estimates of the corresponding parameters (probably due to the small sample size). Using data dependent priors alleviates this issue.

Step 3. Fit a Bayesian model for Equations 8.4 and 8.5 and obtain Markov Chain Monte Carlo (MCMC) draws for all parameters. Preliminary analyses indicated that the chains converge to the posterior by 100,000 iterations. We discarded the first 100,000 iterations, and ran an additional 100,000 iterations to approximate the posterior distribution.

Step 4. Approximate and summarize the posterior distributions of the mediated effects. The first approach to evaluating the size of the mediated effects requires approximating the posterior distributions of these parameters by computing the products $b_{2M}b_{4Y}$ and $b_{3M}b_{4Y}$ using the 100 000 retained draws for these parameters. In order to make inferences about the values of the indirect effects, the posterior distributions need to be summarized using point and interval summaries. Here we use the posterior median instead of the posterior mean because the distribution of the product of two regression coefficients is often asymmetric (Craig, 1936; Lomnicki, 1967). The two options for interval summaries of the posterior are the equal-tail credibility intervals obtained using the alpha/2 and 1-alpha/2 percentiles of the posterior distribution, and the Highest Posterior Density (HPD) intervals which have the property that no value outside of the interval is more probable than values within the interval. Given the potential asymmetry of the posteriors for the indirect effects, we use HPD intervals. The last summary of the posterior is the probability that the mediated effect is of the hypothesized sign (here, positive) computed as the proportion of posterior draws of the mediated effect that are either 0 or positive (as illustrated in Miočević et al., 2017).

Step 5. Test hypotheses that the mediated effects are nonzero. The second approach to evaluating whether the indirect effects are different from 0 requires the specification of hypotheses that evaluate the presence of a mediated effect (akin to the joint significance test in the frequentist framework where the presence of a mediated effect is established if the a-path and b-path in the single mediator model are both significantly different from zero; for more on the logic and statistical properties of the joint significance test, see MacKinnon et al., 2002). A set of four hypotheses of interest, presented in Table 8.3, was defined for the Playskin Lift™dataset presented in this paper. These hypotheses were formulated based on theoretical expectations for the current dataset. For other research questions, the expected signs of the a and b-paths may be different. Because the a-path can be conceptualized in two ways, this set of hypotheses was evaluated using both b_{2M} and b_{3M} as the a-path, while the b-path was conceptualized as b_{4Y}, as shown in the third and fourth columns of Table 8.3.

This set of hypotheses can be used to test the presence of a positive mediated effect. The first hypothesis specifies our main theoretical expectation, namely that both the a-path and the b-path are positive and different from zero. We can compare this hypothesis to its complement, H_{1c}, that says that either the a-path, or the b-path, or both are not positive. This is a generic 'catch-all' alternative hypothesis. By comparing H_1 to H_{1c} we can evaluate whether there is a hypothesized positive mediated effect or not. Additionally, H_2 and H_3 are more precise falsifications of the hypothesized mediated effect under H_1. H_2 specifies that the a-path is negative (as opposed to positive under H_1), without placing any constraints on the b-path. H_3 specifies that the b-path is negative (as opposed to positive in H_1), without placing any constraints on the a-path.

Table 8.2
Mediation hypotheses for the Playskin Lift^{TM} dataset.

Hypothesis	In words	a-path as change in level	a-path as change in trend
H_1: a-path > 0 & b-path > 0	both the a-path and the b-path are positive	H_1: $b_{2M} > 0$ & $b_{4Y} > 0$	H_1: $b_{3M} > 0$ & $b_{4Y} > 0$
H_{1c}: not H_1	either the a-path or the b-path or both are not positive	H_{1c}: not H_1	H_{1c}: not H_1
H_2: a-path < 0	the a-path is in opposite direction (negative)	H_2: $b_{2M} < 0$	H_2: $b_{3M} < 0$
H_3: b-path < 0	the b-path is in opposite direction (negative)	H_3: $b_{4Y} < 0$	H_3: $b_{4Y} < 0$

Bayes factors can be used to compare each pair of these hypotheses to each other and quantify the relative evidence for each hypothesis. The R package bain (Gu et al., 2019) was used to evaluate the above hypotheses. To obtain the Bayes factors, bain requires the sample size and the estimated covariance matrix for the parameters in the hypotheses, which we obtained from the MCMC output in Step 3. The interested reader is referred to the bain manual (Hoijtink, Mulder, Lissa, & Gu, 2019).

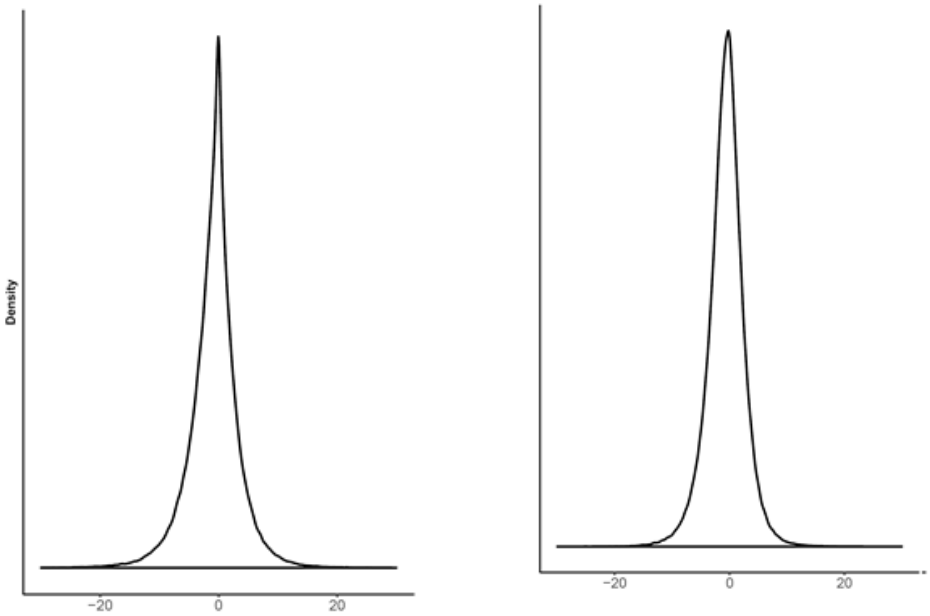
A Bayes factor quantifies the evidence for one hypothesis relative to another. For example, if $BF_{12} = 3$, this means that the data are three times more likely to occur if H_1 is true compared to when H_2 is true. If all pairwise Bayes factors for a set of hypotheses are known, these can be used to update the prior probabilities of the hypotheses to obtain the posterior probabilities. Each hypothesis has a prior probability, that is, the probability that a hypothesis is true before observing the data. Using the posterior probabilities for a set of hypotheses, we can select the best hypothesis from a given set.

8.3 Results

Across all 10 participants, the original study by Babik et al. (2019) found significant improvement of the mean of Grasping and Looking between the baseline and intervention phase. Looking and only unimanual grasping at the object had a significant immediate change at the beginning of the intervention phase. Compared to the time trend in the baseline phase, Grasping had a larger time trend (i.e. rate of improvement) in the intervention phase, but Looking did not have a significantly larger time trend in the intervention phase. Thus there is some evidence for an effect of the independent variable on the dependent variable (path c in the top panel of Figure 8.1), and for an effect of the independent variable on the mediator (path a in the bottom panel of Figure 8.1). The mediation analysis presented below provides additional insights about whether the effect of the intervention on Looking is mediated by improvement in Grasping for one of the participants.

8.3.1 First method: Parameter estimation

The results from Step 4 require evaluating the posterior distribution of the mediated effects $b_{2M}b_{4Y}$ and $b_{3M}b_{4Y}$. The posterior summaries of the mediated effects are presented in Table 8.3 and shown in Figure 8.3. Note that the posterior medians for both mediated effects were negative. The Highest Posterior Density (HPD) intervals for the indirect effect through changes in the level of the mediator, $b_{2M}b_{4Y}$, ranged from −9.885 to 7.881, thus indicating that 0 is among the most probable values for this effect. Furthermore, 42% of the posterior draws were positive, thus indicating that there is 42% probability that the indirect effect through changes in the level of the mediator is positive. The HPD intervals for the indirect effect through changes in the trend of the mediator, $b_{3M}b_{4Y}$, ranged from −7.434 to 5.915, thus indicating that 0 is, once again, among the most probable values for this effect. Furthermore, 42% of the posterior draws were positive, thus indicating that there is 42% probability that the indirect effect through changes in the trend of the mediator is positive. Overall, the posterior summaries suggest that there was no indirect effect of phase change on Looking through changes in level or trend of Grasping. Thus, in this case, no evidence of mediated effect was found.



(a) Indirect effect (level) *(b)* Indirect effect (trend)

Figure 8.3. Plot of posteriors for the mediated effects through the changes in level ($b_{2M}b_{4Y}$) and trend ($b_{3M}b_{4Y}$).

Table 8.3
Posterior summaries of $b_{2M}b_{4Y}$ and $b_{3M}b_{4Y}$.

	$b_{2M}b_{4Y}$	$b_{3M}b_{4Y}$
posterior median	-0.481	-0.526
95% HPD interval	[-9.885, 7.881]	[-7.434, 5.915]
p($ab \geq 0$)	42%	42%

8.3.2 Second method: Hypothesis testing

The results from the Bayesian hypothesis comparison for both representations of the a-path are presented in Table 8.4. H1 has the highest posterior probability of the set of hypotheses for both conceptualizations of the a-path. The differences between the results for the two conceptualizations of the a-path are minimal, and for the sake of brevity we will only discuss the results for the change in level. We find that H_1 is .454/.321 $\approx$ 1.41 times more supported by the data than H_3 and .454/.217 $\approx$ 2.09 times more supported than H_{1c}. There appears to be no evidence for a negative b-path (H_2), since each of the other hypotheses receives at least 24 times more support. There is a slight preference for H_1 relative to its complement H_{1c} and H_3.

Note that the posterior probabilities in Table 8.4 were obtained using equal prior probabilities. That is, all hypotheses received the same prior weight in order to make a fair comparison. The results presented here with equal prior probabilities are in agreement with our expectations. Had we encoded our prior beliefs in subjective prior probabilities and updated those with the evidence from the data, the posterior probability for H1 would be higher.

Table 8.4
Posterior probabilities

	a-path as change in level	a-path as change in trend
H_1	.454	.463
H_{1c}	.217	.215
H_2	.009	< .001
H_3	.321	.322

Note. Probabilities in boldface indicate the hypothesis with the highest probability. These probabilities were obtained with equal prior probabilities.

8.4 Discussion

Identifying mechanisms through which a certain program achieves its effects is extremely important for the identification of the most potent program components and therefore for the conduct of the more evidence-based personalized mental health care (Ng & Weisz, 2015).

In the original SCED study that investigated effectiveness of a Playskin Lift™intervention (Babik et al., 2019) two outcome variables were investigated: Looking at and Grasping for objects. Over the whole group of single-case participants significant improvement of the mean of Grasping and Looking between the baseline and intervention phase was found. However, the theoretical hypothesis underlying Playskin Lift™intervention points to the following: daily exercise with the exoskeletal garment would lead to better grasping ability, and this would, in turn, lead to infant looking more at toys. The testing of this mediating hypothesis was illustrated in the current study using data from one preterm born infant who underwent Playskin Lift™intervention. The methods described in this paper allowed for the computation of the numerical value of the indirect or mediated effect and for testing whether this effect is of the hypothesized sign in SCEDs with two phases (i.e., a baseline phase, A_1, and a treatment phase, B_1) in a single-participant. Bayesian parameter estimation and Bayes Factors are two ways of approaching the same question, however, the results of each approach are interpreted differently and the two approaches may require different numbers of repeated measures of the same participant for optimal performance. We suggest using both approaches in tandem because together they provide more information about the mediated effect(s). In the case of our single participant, no mediated effect of Grasping was found on the Looking efforts of the participant.

We might conclude that for this infant, Playskin Lift™, does not affect looking behavior through changes in grasping behavior, but through some other mechanism, such as increases in parental guidance. In this way individual mechanisms of change could be identified and the most potent therapy techniques that affect changes in these mechanisms. The field of communication disorders could profit from single-case methods in a great way because it is characterized by (i) a great amount of interventions to treat diverse communication problems; (ii) most interventions are seen as evidence-based, as informed by information from group studies; and (iii) large heterogeneity in clients dealing with communication problems.

8.4.1 Limitations

Note that the default coding of the predictors in piecewise regression in the syntax in Appendix A assumes that the phase effect takes place in the first measurement of the second phase. However, change might not be immediate for all therapies, and the syntax needs to be modified to accommodate a different expectation about the timing of the effect. The same is true for the assumed timing of the effect of the mediator: there is a lag of 1 between the mediator and outcome, and this may not be suitable for all processes. Researchers can modify the code we provide to increase the time to the effect, however, in many situations it is very difficult to formulate a prior hypothesis about the appropriate amount of time necessary for changes in the hypothesized mediator to produce changes in the outcome. If a researcher is for instance interested in estimating the effect of the intervention at the third observation point in the intervention phase, then the time can be centered around that observation point. For more information of the influence of centering time on the estimated intervention effect using piecewise regression, see Moeyaert, Ugille, Ferron, Beretvas, & Noortgate (2014).

The Playskin Lift™dataset was limited to only twelve repeated measurements over time. A larger number of observation points is preferred to obtain more certainty in the results. A

simulation study could provide more insight in how much the current sample size affects the results. The sample size also affects the Bayes factor. For any sample size, the Bayes factor has a clear interpretation: the evidence in the data for one hypothesis relative to another. However, it is difficult to say when the evidence is sufficiently strong to rule one hypothesis out. With a small sample size, it is less likely to observe a Bayes factor expressing strong evidence for either hypothesis. Our results showed small Bayes factors and we do not know whether that is because there is indeed only a weak preference for one hypothesis over another, or whether we did not have a sufficient number of observations to obtain stronger evidence.

8.4.2 Future directions

The methods described in the paper have yet to be tested in simulation studies to evaluate the required number of observations per phase for adequate power to detect the mediated effect. Furthermore, future research should develop guidelines and sensitivity analyses for evaluating the timing of the effect of the treatment on the mediator and the effect of the mediator on the outcome.

Data of a single-participant presented in this study was selected from a larger SCED data set, but the same mediation hypothesis can be replicated for the other participants. This data set also used a multiple baseline SCED design (different SCEDs were randomized to different lengths of the baseline A phase). As a consequence, when we replicate our mediation analysis across the other participants, internal and external validity increases. Because frequentist estimates of the regression-based effect sizes have a known sampling distribution, their inverse squared standard error can be used as a weight in meta-analyses. By synthesizing effect sizes across cases and studies, more generalized decisions can be made related to the effectiveness of an intervention, which is a significant contribution to evidence-based practices and policy decisions (Moeyaert et al., 2013a, 2013b, 2014). However, when combining effect sizes across studies, standardization of the outcome score is needed as it is unlikely that exactly the same scale is used across different studies. Future research should extend the methods described in this paper to include standardization, as described by van den Noortgate & Onghena (2007) for frequentist regression-based effect sizes.

8.4.3 Conclusion

This paper illustrated two Bayesian methods for mediation analysis using repeated measures of the potential mediator and outcome of interest from a single participant. The two methods were illustrated using data of a single participant from the Playskin Lift™intervention, and the syntax is provided so researchers can apply the new methods to their data. The new methods have yet to be examined in simulation studies to find out the optimal number of repeated measures required for adequate power to detect the indirect effect in SCEDs. Testing mediators of intervention effects in SCEDs conducted in the field of communication disorders can add valuable information about the mechanisms through which interventions achieve (or do not achieve) the desired effects for a given patient.

Chapter 9

Discussion

In this dissertation practical, philosophical and methodological matters regarding Bayesian informative hypothesis testing have been presented. A summary of the findings and conclusions of the research presented in the dissertation is provided in Section 9.1. Section 9.2 discusses and reflects on the value of these findings and their practical considerations for the field. Section 9.3 concludes this dissertation by discussing potential further research topics and opening the way for a next update on Bayesian informative hypothesis testing.

9.1 A quick summary

Bayesian informative hypothesis testing has been making its appearance in applied social and behavioral research (Mulder & Wagenmakers, 2016). The development of software for Bayesian hypothesis testing (Gu et al., 2019; JASP Team, 2018; Morey & Rouder, 2018) has increased the use of Bayes factors as a tool for hypothesis testing. With increase use, practical problems and theoretical questions present themselves. This dissertation discussed a few of these considerations.

Chapter 2 presented four approaches to determine the sample size when Bayesian informative hypothesis testing is used. Conditional and unconditional error probabilities can be used to plan the appropriate sample size for a study that aims to compare hypotheses with a Bayes factor. Researchers need to define their hypotheses, the expected effect sizes under each of these hypotheses of interest, the desired conditional and unconditional properties of the Bayes factor. Chapter 2 presents four approaches to sample size determination. The level of desired evidence or type of error probability to be controlled determines the required sample size. How this error level should be determined depends on the desired conclusion. This implies there is no *one method fits all* to determine sample size. The R package `BayesianPower` Klaassen (2019) presented in Chapter 6 enables applies researchers to compute the required sample size for their Bayesian informative hypothesis test.

Chapters 3 and 4 illustrate the importance of matching the research question to the analysis. Many hypotheses tests evaluate the presence of a population effect size, while the

conclusions and research questions concern individial effects. An effect at the population level might not hold at the individual level. Chapter 3 presents a method that allows for the analysis of individual effects and aggregating that information. This method differs from updating or sequential analysis, which means that hypotheses can be evaluated or parameters estimated repeatedly as more data becomes available. Aggregating or synthesizing evidence implies that the information from multiple sources is combined to draw conclusions about the collection of datasets. This distinction between updating and evidence synthesis is an important distinction emphasizes the importance of awareness of the question of interest. Chapter 4 presents a step by step description of how to execute such an analysis, and Chapter 6 presents the R Shiny application that makes these methods accessible.

Chapter 5 considers the importance of specifying prior probabilities to better enable knowledge updating. Any individual research project contributes only one step of the updating cycle. It is important to always consider the larger cycle, so that the appropriate information can be provided. While the evidence might be the relevant output of a project, in order to enable other to use and evaluate your output prior probabilities are required. Without prior probabilities, readers cannot evaluate the obtained evidence properly. Therefore, it is important to specify the prior probabilities chosen and the rationale behind them. Chapter 5 defines and discusses the role of prior probabilities, and presents and evaluates an elicitation procedure for prior probabilities.

Chapter 7 and 8 present two applied projects where informative hypotheses were evaluated, demonstrating the wide range of applications of informative hypotheses. In Chapter 7 an experimental repeated measures dataset is presented. The dataset is modeled by means of a hierarchical model. The theoretical expecations are translated in informatve hypotheses and evaluated by means of Bayes factors. Chapter 8 proposes a method to evaluate the presence and direction of individual mediated effects. An illustration shows how individual informative hypotheses can be used to evaluate expectations This illustration shows the flexibility of Bayesian informative hypothesis testing. The hypotheses use combinations of constraints to match the theoretical expectation of the mediated effect. Furthermore, the analysis is executed at the level of the individual effects rather than the population effects.

9.2 Bayesian informative hypothesis evaluation

This dissertation presents a wide variety of applications, considerations and implications of Bayesian informative hypothesis testing. How do these chapters contribute to updating our knowledge of the value of Bayesian informative hypothesis testing?

9.2.1 Informative hypotheses

The value of *informative* hypotheses can be defined in different ways. First and foremost, it is valuable because it specifies the exact interest and question of a researcher. Especially in experimental research where different conditions or treatments are compared, researchers might have expectations about the direction or (relative) size of effects. This argument is much cited in relation to the criticism of NHST. Informative hypotheses allow for the evaluation of expectations and compare specific theories, while null hypotheses cannot be

confirmed, nor do they seem likely possible true representations. A second valuable aspect of informative hypotheses that is presented throughout this dissertation is the value of adding knowledge to an analysis. The specification of a set of informative hypotheses might be more challenging or difficult than applying a default null and alternative hypothesis. However, the value in this difficulty is that the answer is much more interpretable and the scope of generalizations of such a hypothesis test are more tangible. By taking the effort to define the exact hypotheses, a researcher becomes more aware of the conclusions that are and importantly, that are not possible to draw from the analysis (e.g. Chapter 3 and 4. This awareness might cause a more specific comparison to be made (that may result in smaller required sample sizes, see Chapter 2). Importantly, using an informative hypothesis helps to express the question of interest.

9.2.2 Bayesian hypothesis evaluation

The informative hypotheses in this dissertation are considered in a *Bayesian* context. Bayesian hypothesis testing is valuable because multiple theoretical expectations can be compared to each other. Additionally, it is straightforward to update knowledge sequentially. The Bayesian framework invites a researcher to think about prior distributions and prior probabilities. It offers many possibilities to model the analysis exactly to the question or balance the evidence and error rates. The necessity of making choices and justifying them could be seen as a burden, but really provides a plentitude of opportunities. Chapter 2 showed how choices have to be made about the desired properties of a Bayesian hypothesis test. Chapters 3 and 4 showed the importance of asking the right question of interest. Chapter 5 demonstrated how personal justification makes results more interpretable. Justification is the most important key for any of these issues.

9.2.3 Updating knowledge

The final concept evaluated in this dissertation is the *updating* of knowledge. In any single research, the focus is on the results of that one study. It is tempting to put the responsibility of updating at the next researcher. However, updating knowledge is a shared process. Updating is not exclusively for meta-analysts or systematic reviewers. Updating is something everyone can contribute to by presenting their evidence such that it is ready to be used and updated. It is important to formulate how knowledge can be updated on the basis of a study. This invites readers to reflect more on this evidence and incorporate it in their future research. In other words, the research does depend on each other, we just do not make it explicit generally how this dependency can be traced back. The outcome of this latest update regarding Bayesian informative hypothesis testing is that it comes with many more challenges, but offers even more opportunities.

9.2.4 Practical considerations

In order to use Bayesian informative hypothesis evaluation, it is important to be aware of the research question and goal of the research. This may seem obvious: of course researchers think about what their research question is. Any academic paper presents a

research question and any writing course or guidelines will tell you that there is no message without a question. Statistical analyses might be chosen for their familiarity, even though they might not match the research question adequately. That is, if you have been thought to test null hypotheses, that is what you know, and those are the questions you ask. There are many steps in statistical analyses that are important to question, as this dissertation shows. If a researcher is designing an experiment and wants to determine the appropriate sample size, they need to be aware of the question of interest. As Chapter 2 has shown, this evokes many subsequent questions: how are the populations and effect sizes defined under each considered hypothesis and what errors are you willing to make? Chapters 3 and 4 introduce the questioning about who a hypothesis should apply to. Is the research interest in individuals or in an average effect? Finally, Chapter 5 discusses the distinction in answering a research question: is the goal to describe how the current state of knowledge can be changed by the findings or to describe the current state of knowledge? All these matters depend on the goal of the research and should be considered as part of the definition of a research question.

A second critical aspect of applying Bayesian informative hypothesis evaluation in practice, is the current state of knowledge. Again, this might seem like an obvious aspect of research. Literature reviews are directed at defining a gap in knowledge and describing the expectations for a current research. However, as discussed at various points in this dissertation, it is important to included these expectations in the statistical analyses. If there is a mismatch between the expectations and the analysis, the conclusions from this analysis too do not match the initial question. Practically, there are many aspects in a research project where knowledge is available and used, either explicitely or implicitely. Similarly, the knowledge about the specification of hypotheses, the exclusion of hypotheses, the choice in the set of interest, the prior probability of hypotheses. All these elements of knowledge are generally not described in a paper. This dissertation contributes to the more explicit description of choices made, and alternatives that were rejected before coming to the chosen analysis plan. All these pieces of information are a valuable part to the output that is interpreted by others that know nothing of these choices.

9.3 Concluding remarks

9.3.1 Future directions

With raising awareness to opportunities of Bayesian informative hypothesis testing and the important questions to be asking, many open ends and further research opportunities can be identified. In consultations about informative hypothesis testing, one of the most frequent questions is: which and how many hypotheses should I compare? Particularly the difference between using a complement or an unconstrained hypothesis to evaluate the value of a single hypothesis is an often-raised question. Chapter 2 briefly touches upon this issue but no thorough investigation of the impacts of this difference have been made. Evidently, the meaning of the conclusions coming from a comparison of an informative hypothesis to its complement are different from those from a comparison to the unconstrained hypothesis. Particularly, when only inequality constraints are considered, the Bayes factor against an unconstrained hypothesis has an upper limit that is meaningful. Rather than ranging

between − inf and inf the evidence in favor of an inequality constrained hypothesis relative to the unconstrained can be only as high as the inverse of its complexity. This is true because it is nested in the unconstrained hypothesis: if $H_i : \theta_1 > \theta_2$ is true, $H_u : \theta_1, \theta_2$ is also true. The only difference is that H_u describes a greater range of possibilities and therefore has a lower density, which is a function of the complexity of H_i. For the discussion above only two hypotheses were considered. The matter gets more complicated if more hypotheses are evaluated, or if the interest is in parts of an informative hypothesis too. The size of the set of hypotheses, the inclusions of one or multiple complementary hypotheses, or the consideration of multiple 'subset' hypotheses all influence the (un)conditional error probabilities. If a hypothesis does not hold *for everybody*, it does not imply that it holds for nobody. Further analysis of the individual datasets might show the presence of subgroups in terms of the theoretical constraints.

9.3.2 The latest update

Bayesian informative hypothesis tests can be tailored to fit your research question exactly. Many options and choices have to be considered before the actual analysis can be done. Rather than avoid these choices or aim for golden standards and default options, we should investigate the flexibility it offers us. Bayesian informative hypothesis testing challenges us to be specific about the question we ask and where they come from. The questions we ask today are different from the questions we asked a year ago. Those past questions are the steps that have led us to where we are today.

Bayesian informative hypothesis testing invites us to think about our theories and be explicit about them. Had it been my mother who could whistle and my father who could not, I might have developed a theory that only mothers can whistle. By considering all the options in an analysis explicitely, we know better what we can and cannot claim based on the outcomes. My observations as a six-year old rejected the theory that gender determines whistling abilities. I never bothered to investigate whether there is a gender difference in whistling abilities at the population level. Bayesian informative hypothesis testing allows us to update our knowledge, to make mistakes and learn from them. Update your knowledge. Do not be afraid of being wrong, after all: it is all just one more step, leading to the next one. Even though I was convinced of my personal theory, I adjusted my theory and continued to learn more from there. Learning does not stop when I have learned to whistle a symphony, or when I've learned how to perfect my technique. Every question is informed by previous answers and will inform next wonderings. What will the next update be?

Chapter 10

Appendices

10.1 Chapter 2. The power of informative hypotheses

Appendix 1. Numerical characteristics of tables

This appendix illustrates some numerical characteristics of Tables 2.4–2.17 by means of examples.

First, in all tables the required sample size increases if the error probability or Indecision probability decreases, or if B increases. Put differently, the more certainty is required for the conclusion, the larger the sample size should be. If the violation size under $H_{i'}$ increases, the required sample size decreases. Hypotheses with larger violations are more distinctly different from H_i: datasets generated under H_i will less often result in a decision in favour of $H_{i'}$, and vice versa, compared to small violations.

Second, if K increases, a larger sample size is required. If K increases, but d_{H_i} is constant, the differences between pair of means decreases. For example, if $d_{H_i} = .5$, the difference between each pair of means is .5 for $K = 2$, .25 for $K = 3$, and .167 for $K = 4$. If differences between means are smaller, it is more likely that the means of a sample will not adhere to the population from which they were sampled, thus, a larger sample size is required.

Furthermore, in three situations, some symmetry is visible in the tables. First, if d_{H_i} and d_{H_c} are equal with $K = 2$, H_i and H_c are exchangeable. Therefore, the Type i and Type c error probabilities are equal. Since the Decision error probability is their average, it holds that this is equal to both the Type i and Type c error probability. Thus, the required sample size is independent of the type of error probability controlled. For example, as can be seen in Table 2.4, for $K = 2$, $d_{H_i} = .2$, and a critical value for the error probabilities of .05, the group sample size is 132, whether the Decision error, Type i, or Type c error probability is controlled.

Second, if d_{H_i} and $d_{H_{i'}}$ are equal, H_i and $H_{i'}$ are exchangeable as well, and thus the Type i and Type i' error probabilities are equal. Thus, the sample size is independent of the type of error controlled. As can be seen in Table 2.5, for $K = 3$, $d_{H_i} = d_{H_{i'}} = .5$, and $H_{i'}$ with a

large violation size, and a critical value for the error probability of .05, the sample size is 22, whether the Decision error, Type i, or Type c error probability is controlled.

Third, d_{H_i} and $d_{H_{i'}}$ can be unequal in two ways: $d_{H_i} = .2$ and $d_{H_{i'}} = .5$, or $d_{H_i} = .5$ and $d_{H_{i'}} = .2$. The Type i error probability for $d_{H_i} = .2$ and $d_{H_{i'}} = .5$ will be the same as the Type i' error probability for $d_{H_i} = .5$ and $d_{H_{i'}} = .2$, and vice versa. Thus, the Decision error probability will be exchangeable in these situations. Therefore, the required sample size to control the Decision error probability will be the same whether the first or the second set of effect sizes is used. As can be seen in Table 2.5, for $K = 4$, $H_{i'}$ with a small violation, and a critical value for the Decision error probability of .1, for $d_{H_i} = .5$ and $d_{H_{i'}} = .2$ or $d_{H_i} = .2$ and $d_{H_{i'}} = .5$, the sample size is 338.

Note that examples of the three situations described above exist where the sample sizes are not exactly equal, but about equal. This is caused by the sampling variation of he simulation procedure presented in Section 2.5.4.

Contrary to expectations, the required sample size when comparing H_i with $H_{i'}$ is not always smaller than when comparing H_i with H_c. As an example, a repetition of Example 1.3 from Section 2.6.1 follows:

Example 1.3 Suppose a researcher wants to evaluate H_i with H_c, with $K = 4$. The researcher wants to control the Type i error probability at .025. He specifies $d_{H_i} = .2$, $d_{H_c} = .2$. As can be seen in Table 2.4, the required sample size is 443. Suppose this researcher did not consider H_c, but $H_{i'}$. As can be seen in Table 2.5, the required sample size is larger than $1,000$ if $H_{i'}$ was specified with a small violation of H_i, the sample size is 575 with a medium violation, and 169 with a large violation.

In this example, the required sample size to control the Type i error comparing H_i with H_c is smaller than the sample size when comparing H_i with $H_{i'}$, only for medium and small violations of $H_{i'}$, but not for large violations of $H_{i'}$. This can be explained best by means of an example: Let us consider $K = 4$, such that $c_i = c_{i'} = 1/24$, and $c_c = 23/24$, then,

$$BF_{ic} = \frac{f_i/c_i}{f_c/c_c} = \frac{f_i/\frac{1}{24}}{f_c/\frac{23}{24}} = \frac{f_i}{f_c/23} = f_i \cdot 23\frac{1}{f_c}$$

$$BF_{ii'} = \frac{f_i/c_i}{f_{i'}/c_{i'}} = \frac{f_i/\frac{1}{24}}{f_{i'}/\frac{1}{24}} = \frac{f_i}{f_{i'}} = f_i \cdot \frac{1}{f_{i'}}$$

Thus, if the fit of the data to H_c is five times larger than the fit of the data to $H_{i'}$, BF_{ic} and $BF_{ii'}$ will be equal. If the fit of the data to H_c is less than five times the fit to $H_{i'}$, BF_{ic} will be larger than $BF_{ii'}$, and if it is more than five times the fit to $H_{i'}$, BF_{ic} will be smaller than $BF_{ii'}$. If Bayes factors based on population in which H_i is true are larger, the Type i error probability becomes smaller.

In Example 1.3, only the Type i error is considered, which means that only data generated under H_i is considered. Thus, f_i will be the same for BF_{ic} and for $BF_{ii'}$. However, the fit to H_c or to $H_{i'}$ is also taken into account in the computation of the Bayes factor. Any part of the posterior distribution that does not fit to H_i, will fit to H_c. Since $H_{i'}$ is a subset of H_c, the fit to $H_{i'}$ will always be smaller than that to H_c.

More often than not, the fit of data generated under H_i will be larger to $H_{i'_{\text{small}}}$, than to

$H_{i'_{\text{large}}}$. If data are generated based on a population in which H_i holds, it is more likely to obtain a sample that adheres to $H_{i'_{\text{small}}}$ than to $H_{i'_{\text{large}}}$. In general, using data simulated from a population in which H_i is true, Bayes factors concerning $H_{i'_{\text{large}}}$ will be larger than those concerning $H_{i'_{\text{small}}}$, and thus the Type i error probability will be smaller for $H_{i'_{\text{large}}}$ than for $H_{i'_{\text{small}}}$.

This explains the differences in sample size for the different violation sizes under $H_{i'}$. Applying this to Example 1.3, it appears that for a large violation under $H_{i'}$, the fit to H_c is more than 23 times larger than to $H_{i'}$. thus the required sample size for a large violation is smaller than that for H_c.

10.2 Chapter 3. All for one or some for all? Evaluating informative hypotheses using multiple $N = 1$ studies

10.2.1 Appendix 1. Computation of fit and complexity through decomposition

In order to compute the Bayes factor that expresses the support in favor of H_m:

$$H_m : R_m\boldsymbol{\pi} > 0 \tag{10.1}$$

against the unconstrained hypothesis H_u, the complexity and fit of H_m should be computed[1].

Complexity and fit can be determined by taking samples from the unconstrained prior and posterior distribution respectively. A common approach is to take Q samples, and determine what proportion of the samples is in agreement with H_m, such that

$$\begin{aligned} f_m &= \int_{\boldsymbol{\pi}\in H_m} g(\boldsymbol{\pi}|\mathbf{x}, H_u)\delta\boldsymbol{\pi} \\ &\approx \frac{1}{Q}\sum_{q=1}^{Q} I_{\boldsymbol{\pi}^q\in H_m} \end{aligned} \tag{10.2}$$

where $\boldsymbol{\pi}^q$ is the qth sample from the unconstrained posterior and $I_{\boldsymbol{\pi}^q\in H_m} = 1$ if $\boldsymbol{\pi}^q$ is in agreement with H_m, and 0 otherwise. The complexity can be computed analogously, with the difference that samples are taken from the prior distribution.

If H_m concerns the ordering of 8 parameters, the complexity can be derived analytically and is $1/8! = 1/40,320$. Using $Q = 100,000$ samples from the unconstrained prior only 2 or 3 samples of $\boldsymbol{\pi}$ are expected to adhere to the constraints under H_m. This implies that the estimate of f_m is very unstable. To obtain stable estimates impossibly huge samples are needed. Similarly, the fit of a hypothesis with 8 parameters might be too small to accurately approximate using $100,000$ samples. One solution is to increase the number of samples which increases the computational time. Mulder et al. (2012) present another solution that makes use of a decomposition of the complexity and fit. This procedure determines

[1]Note that for notational simplification the superscript i is dropped from the hypotheses, Bayes factors, and parameters in this appendix.

decomposed fit and complexity for each constraint in a hypothesis. Equation 10.3 shows how the probability that all constraints hold, given H_u and the data $\mathbf{x}$, can be rewritten as a product of decomposed probabilities:

$$\begin{aligned} P(\mathbf{R}_m\boldsymbol{\pi} > 0|H_u, \mathbf{x}) &= \prod_{k=1}^{K} P(\mathbf{R}_m^{(k)}\boldsymbol{\pi} > 0|H_u, \mathbf{x}, \mathbf{R}_m^{(1:k-1)}) \\ &= \prod_{k=1}^{K} f_m^{(k)} \\ &\approx \prod_{k=1}^{K} \frac{1}{Q}\sum_{q=1}^{Q} I_{\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q>0}, \end{aligned} \tag{10.3}$$

where K is the number of constraints in hypothesis m, $\mathbf{R}_m^{(k)}$ is the k^{th} row of $\mathbf{R}_m$, $\mathbf{R}_m^{(1:k-1)}$ are the first $k-1$ rows of $\mathbf{R}_m$, $f_m^{(k)}$ is the decomposed fit for the k^{th} constraint, the indicator function $I_{\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q>0} = 1$ if $\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q > 0$ and 0 otherwise and $\boldsymbol{\pi}^q$ is sampled from $g(\boldsymbol{\pi}|H_u, \mathbf{x}, \mathbf{R}_m^{(1:k-1)})$.

Since each $f_m^{(k)}$ is only defined by one constraint, it is never a small value and can be estimated with relatively few samples. The R Shiny application *OneForAll* belonging to this paper uses $Q = 10,000$. By multiplying the decomposed fit components similar to Equation 10.3 the total fit can be obtained accurately.

The complexity can be derived analogously:

$$\begin{aligned} P(\mathbf{R}_m\boldsymbol{\pi} > 0|H_u) &= \prod_{k=1}^{K} P(\mathbf{R}_m^{(k)}\boldsymbol{\pi} > 0|H_u, \mathbf{R}_m^{(1:k-1)}) \\ &= \prod_{k=1}^{K} c_m^{(k)} \\ &\approx \prod_{k=1}^{K} \frac{1}{Q}\sum_{q=1}^{Q} I_{\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q>0}, \end{aligned} \tag{10.4}$$

where $c_m^{(k)}$ is the decomposed complexity conditional for the k^{th} constraint and $\boldsymbol{\pi}^q$ is sampled from $h(\boldsymbol{\pi}|H_u, \mathbf{R}_m^{(1:k-1)})$.

10.3 Chapter 7. Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence

10.3.1 Appendix 1. Model statement in JAGS

```
model{
```

```
#model
  for(i in 1:N){
    y[i]~dnorm(y.hat[i], tau.e)

    y.hat[i] <- b00 + b0c*Condition[i] + b0v*Valence[i] +
                b0a*Arousal[i] + u[person[i]]
  }
  # prior distributions
  b00 ~ dnorm(0, .000001)
  b0c ~ dnorm(0, .000001)
  b0v ~ dnorm(0, .000001)
  b0a ~ dnorm(0, .000001)
  tau.e ~ dgamma(.01, .01)

# random effect
  for(j in 1:J){
    u[j] ~ dnorm(0, tau.u)
  }

  tau.u ~ dgamma(.01, .01)

  sigma.e <- 1/tau.e
  sigma.u <- 1/tau.u
}
```

10.3.2 Appendix 2. Model statement in JAGS for extended model

```
model{
  for(i in 1:N){
# model
    y[i]~dnorm(y.hat[i], tau.e)
    y.hat[i] <- b00 +
      b0c*Condition[i] +
      b0v*Valence[i] +
      b0a*Arousal[i] +
      bm[1]*movie.f20[i] +
      bm[2]*movie.f2[i] +
      bm[3]*movie.f3[i] +
      bm[4]*movie.f4[i] +
      bm[5]*movie.f5[i] +
      bm[6]*movie.f6[i] +
      bm[7]*movie.f7[i] +
      bm[8]*movie.f8[i] +
      bm[9]*movie.f9[i] +
      bm[10]*movie.f10[i] +
      bm[11]*movie.f11[i] +
      bm[12]*movie.f12[i] +
      bm[13]*movie.f13[i] +
      bm[14]*movie.f14[i] +
      bm[15]*movie.f15[i] +
      bm[16]*movie.f16[i] +
      bm[17]*movie.f17[i] +
      bm[18]*movie.f18[i] +
      bm[19]*movie.f19[i] +
      dm[1]*movie.i20[i] +
      dm[2]*movie.i2[i] +
      dm[3]*movie.i3[i] +
      dm[4]*movie.i4[i] +
      dm[5]*movie.i5[i] +
      dm[6]*movie.i6[i] +
      dm[7]*movie.i7[i] +
      dm[8]*movie.i8[i] +
      dm[9]*movie.i9[i] +
      dm[10]*movie.i10[i] +
      dm[11]*movie.i11[i] +
      dm[12]*movie.i12[i] +
      dm[13]*movie.i13[i] +
      dm[14]*movie.i14[i] +
      dm[15]*movie.i15[i] +
      dm[16]*movie.i16[i] +
      dm[17]*movie.i17[i] +
      dm[18]*movie.i18[i] +
```

```
      dm[19]*movie.i19[i]
      + u[person[i]]
  }
#priors
  b00 ~ dnorm(0, .000001)
  b0c ~ dnorm(0, .000001)
  b0v ~ dnorm(0, .000001)
  b0a ~ dnorm(0, .000001)
  for(j in 1:19){
    dm[j] ~ dnorm(0, .000001)
    bm[j] ~ dnorm(0, .000001)
  }
  tau.e ~ dgamma(.01, .01)

#random effect
  for(j in 1:J){
    u[j] ~ dnorm(0, tau.u)
  }
  tau.u ~ dgamma(.01, .01)
  sigma.e <- 1/tau.e
  sigma.u <- 1/tau.u
}
```

10.3.3 Appendix 3. Supplementary tables extended model

Table 10.1
Parameter estimates

Parameter	HPD Estimate	95% CI	Standard error	Std. coefficient
b_{00}	0.496	[-0.229 : 1.125]	.345	1.506
b_{0a}	0.022	[-0.004 : 0.047]	.013	0.097
b_{0c}	-0.469	[-0.863 : 0.055]	.214	-0.797
b_{0v}	0.028	[0.002 : 0.055]	.013	0.103
τ_e	5.933	[5.238 : 6.675]	.367	1.571
τ_u	14.605	[7.374 : 25.103]	4.562	3.936

Highest posterior density parameters estimates obtained from the Bayesian analysis, with a 95% Credible Interval, standard error and standardized parameter value. b_{00} denotes the intercept, b_{0a}, b_{0c} and b_{0v} the regression coefficient for arousal, condition and valence, respectively, τ_e denotes the residual variance and τ_u the individual intercept variance.

Table 10.2
Bayes factors

Sample size	29	180	380	580
H_1 versus H_{1c}	1.23	3.07	4.46	5.51
H_1 versus H_2	10.08	25.11	36.48	45.07

Bayes factors expressing the relative evidence in the data for H1 versus H1c (top row) or H2 (bottom row) for effective sample sizes 29, 180, 380 and 580. Bayes factors for the unstandardized analysis are presented here. Bayes factors are similar for the standardized analysis.

10.4 Chapter 8. Using Bayesian methods to test mediators of intervention outcomes in Single case experimental designs (SCEDs)

10.4.1 Appendix 1. R script for analyses

```
# This code allows for obtaining results of a Bayesian mediation
# analysis using data from a single participant.
# The results include plots of the posteriors, and point and interval
# summaries of the mediated effect conceptualized as the change in
# the level of the outcome variable following a change in level of
# the mediator and the mediated effect conceptualized as the change
# in the level of the outcome variable following a change in the
# trend of the mediator. The results also include Bayes Factors
# obtained using informative hypothesis testing to evaluate the
# relative support for the hypotheses that the effect # of treatment
# on the mediator and the effect of the mediator on the outcome
# are consistent with theory.
# Make sure you have JAGS installed before starting the analysis.
# You can install JAGS from the following website:
# https://sourceforge.net/projects/mcmc-jags/files/

###############  Installing R packages needed for the analyses #####
install.packages(c("bain", "readr", "rjags", “coda”,
                   "ggplot2", “gridExtra”))

############################ Data information ######################
# The data set needs to contain the following variables
# (with the same names):
# Time = Measurement occasion
# Phase = 0 if phase A, 1 if phase B
# Time1 = Measurement occasion - 1
# Time2 = 0 at the first observation of phase B (computed as Time -
# occasion at the start of phase B)
```

```
# phase_time2 = 0 in phase A and at the first occasion of phase B,
# count starting at 1 from the second observation in phase B
# ScoreM = scores on the mediator at each measurement occasion
# ScoreY = scores on the outcome at each measurement occasion
# Tmed = scores on the mediator with lag 1
# Tout = scores on the outcome with lag 1

################################## Data import ######################
library(readr)
Data_SCED_mediation <-
    # Replace the .csv file name with the name of your data set
    # Make sure your working directory contains the data file
read_csv("Data_SCED_mediation.csv")

######################### Step 1 Frequentist estimates #############
### This part of the code should not be changed
# Unstandardized piecewise regression frequentist estimates for
# phases 1 & 2
reg1 <- lm(ScoreM ~ Time1 + Phase + phase_time2 + Tmed,
           Data_SCED_mediation)
reg2 <- lm(ScoreY ~ Time1 + Phase + phase_time2 + Tmed + Tout,
Data_SCED_mediation)

######################### Step 2 Data dependent priors #############
# Extract the coefficient estimates to be used in the data
# dependent priors
int.mean.m <- coefficients(reg1)[1]
time.mean.m <- coefficients(reg1)[2]
phase.mean.m <- coefficients(reg1)[3]
phasetime.mean.m <-coefficients(reg1)[4]
int.mean.y <- coefficients(reg2)[1]
time.mean.y <- coefficients(reg2)[2]
phase.mean.y <- coefficients(reg2)[3]
phasetime.mean.y <- coefficients(reg2)[4]
tmed.y <- coefficients(reg2)[5]
tout.y <- coefficients(reg2)[6]

################### Step 3 Run Rjags for Bayesian estimates ########
library(rjags)
N<-dim(Data_SCED_mediation)[1]

################## Model definition
modelstring <- as.character("
model{
##### Priors #####
# prior for the intercept in the equation predicting M
beta.0M ~ dnorm(int.mean.m, .001);
# prior for b1
```

```
beta.1M ~ dnorm(time.mean.m, .001);
# prior for b2
beta.2M ~ dnorm(phase.mean.m, .001);
# prior for b3
beta.3M ~ dnorm(phasetime.mean.m, .001);
# prior for b4M
beta.4M ~ dnorm(tmed.y, .001);
# prior for the error precision of M
tau.eM ~ dgamma(.5, .5);
# prior for the intercept of Y
beta.0Y ~ dnorm(int.mean.y, .001);
# prior for b1Y
beta.1Y ~ dnorm(time.mean.y, .001);
# prior for b2Y
beta.2Y ~ dnorm(phase.mean.y, .001);
# prior for b3Y
beta.3Y ~ dnorm(phasetime.mean.y, .001);
# prior for b4Y
beta.4Y ~ dnorm(tmed.y, .001);
# prior for b5Y
beta.5Y ~ dnorm(tout.y, .001);
# prior for the error precision of Y
tau.eY ~ dgamma(.5, .5);
# priors for the missing data
Phase[N] ~ dnorm(0, .000001);
Time1[N]~ dnorm(0, .000001);
Tmed[1]~ dnorm(0, .000001);
Tout[1]~ dnorm(0, .000001);
phase_time2[N]~ dnorm(0, .000001);
##### Conditional probability of the data #####
# The regression model
for(i in 1:N){
m.prime[i] <- beta.0M + beta.1M*Time1[i] + beta.2M*Phase[i] +
              beta.3M*phase_time2[i] + beta.4M*Tmed[i]
y.prime[i] <- beta.0Y + beta.1Y*Time1[i] + beta.2Y*Phase[i] +
              beta.3Y*phase_time2[i] + beta.4Y*Tmed[i] +
              beta.5Y*Tout[i]
# conditional distribution of M
ScoreM[i] ~ dnorm(m.prime[i], tau.eM);
# conditional distribution of Y
ScoreY[i] ~ dnorm(y.prime[i], tau.eY);
} # closes the regression model
}
")
model.file.name <- "Linear Regression.txt"
write(x = modelstring, file = model.file.name, append = FALSE)

################### Run the Bayesian piecewise regression in rjags
```

```
# set seed for replicable MCMC samples
set.seed(123)
jags <- jags.model('Linear Regression.txt',
data = list('Time1' = Data_SCED_mediation$Time1,
'Phase' = Data_SCED_mediation$Phase,
'phase_time2' = Data_SCED_mediation$phase_time2,
'ScoreM'= Data_SCED_mediation$ScoreM,
'ScoreY'= Data_SCED_mediation$ScoreY,
'Tmed' = Data_SCED_mediation$Tmed,
'Tout' = Data_SCED_mediation$Tout,
'int.mean.m' = int.mean.m,
'time.mean.m' = time.mean.m,
'phase.mean.m' = phase.mean.m,
'phasetime.mean.m' = phasetime.mean.m,
'int.mean.y' = int.mean.y,
'time.mean.y' = time.mean.y,
'phase.mean.y' = phase.mean.y,
'phasetime.mean.y' = phasetime.mean.y,
'tmed.y' = tmed.y,
'tout.y' = tout.y,
'N' = N),
n.chains = 3)
out <- coda.samples(jags, variable.names = c("beta.0M",
                                             "beta.1M",
                                             "beta.2M",
                                             "beta.3M",
                                             "beta.4M",
                                             "tau.eM",
                                             "beta.0Y",
                                             "beta.1Y",
                                             "beta.2Y",
                                             "beta.3Y",
                                             "beta.4Y",
                                             "beta.5Y",
# If missing values are of interest, remove the comments in the
# lines below and replace the number “13” with the number of
# observations in your data set plus 1.
                                             # "Phase[13]",
                                             # "Time1[13]",
                                             # "Tmed[1]",
                                             # "Tout[1]",
                                             # "phase_time2[13]",
                                             "tau.eY"),
         n.iter = 100000)
summary(out)

###### Diagnose convergence
library(coda)
```

```
model.as.mcmc.list <- as.mcmc.list(out)
gelman.diag(model.as.mcmc.list)
gelman.plot(model.as.mcmc.list)
plot(model.as.mcmc.list, trace = TRUE, density = FALSE)

###### Running additional iterations
out2 <- coda.samples(jags, variable.names = c("beta.0M",
                                              "beta.1M",
                                               "beta.2M",
                                               "beta.3M",
                                               "beta.4M",
                                               "tau.eM",
                                               "beta.0Y",
                                               "beta.1Y",
                                               "beta.2Y",
                                               "beta.3Y",
                                               "beta.4Y",
                                               "beta.5Y",
                                             # "Phase[13]",
                                             # "Time1[13]",
                                             # "Tmed[1]",
                                             # "Tout[1]",
                                             # "phase_time2[13]",
                                               "tau.eY"),
                                            n.iter=100000)
###### Diagnose convergence
model.as.mcmc.list <- as.mcmc.list(out2)
gelman.diag(model.as.mcmc.list)
gelman.plot(model.as.mcmc.list)
plot(model.as.mcmc.list, trace = TRUE, density = FALSE)
summary(out2)

###### Collect final draws
draws <- as.mcmc(do.call(rbind,model.as.mcmc.list))
draws <- as.data.frame(draws)

#### Step 4 Posterior distributions of the mediated effects   ######
##### Draws for mediated effect(s)
# Option 1 for the a-path: change in level at the start of phase B
# (beta.2M)
# (beta.3M)
ablevel <- draws$beta.2M * draws$beta.4Y
abtrend <- draws$beta.3M * draws$beta.4Y

##### Plot the posterors of the mediated effects
library(ggplot2)
level <- ggplot(as.data.frame(ablevel)) +
geom_density(aes(x = ablevel), size = 1) +
```

```
scale_x_continuous(limits=c(-30,30)) +
labs(x = NULL) +
theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
panel.background = element_blank(),
axis.line = element_line(colour = "black"),
axis.text.x  = element_text(size = 14),
axis.text.y = element_blank(),
axis.title.y = element_text(size = 14),
axis.ticks.y =  element_blank()) +
ylab("density") +
labs(title = "Indirect effect (level)", x = " ", y = "Density") +
theme(plot.title = element_text(color = "black",
size = 14, face = "bold", hjust = 0.5),
axis.title.x = element_text(color="black", size=12, face="bold"),
axis.title.y = element_text(color="black", size=12, face="bold")) +
scale_color_grey()
trend <- ggplot(as.data.frame(abtrend)) +
geom_density(aes(x = abtrend), size=1) +
scale_x_continuous(limits=c(-30,30)) +
labs(x = NULL) +
theme(panel.grid.major = element_blank(),
panel.grid.minor = element_blank(),
            panel.background = element_blank(),
            axis.line = element_line(colour="black"),
            axis.text.x = element_text(size=14),
            axis.text.y = element_blank(),
            axis.title.y = element_text(size=14),
            axis.ticks.y =  element_blank()) +
labs(title = "Indirect effect (trend)", x = " ", y = " ") +
theme(plot.title = element_text(color = "black",
size = 14, face = "bold", hjust = 0.5),
axis.title.x = element_text(color="black", size=12, face="bold"),
axis.title.y = element_text(color="black", size=12, face="bold")) +
scale_color_grey()

##### Make a panel containing both plots
library(gridExtra)
grid.arrange(level, trend, nrow = 1)

##### Obtain mean, medians, and quantiles for the mediated effect(s)
summary.stats.level <- summary(as.mcmc(ablevel))
summary.stats.trend <- summary(as.mcmc(abtrend))
summary.stats.level
summary.stats.trend

#### Obtain highest posterior density (HPD) intervals for the
# mediated effects
```

```
HPD.interval.level <- HPDinterval(as.mcmc(ablevel), prob=.95)
HPD.interval.trend <- HPDinterval(as.mcmc(abtrend), prob=.95)
HPD.interval.level
HPD.interval.trend

#################### Step 5 Bayesian hypothesis testing ###########
library(bain)
# set seed for hypothesis testing
set.seed(1234)
# Collect estimates and covariance matrix
estimate <- colMeans(draws)
cov1 <- cov(draws)

###### Note
### This syntax tests the hypotheses that both the a and b paths are
### positive. This part of the syntax will need to be changed if your
### hypotheses are that one or both paths are negative.

### Option 1 for the a-path: change in level at the start of phase B
### (beta.2M)
results.level <- bain(estimate,
"beta.2M > 0 & beta.4Y >0;
beta.2M < 0;
beta.4Y < 0",
n = N, Sigma = cov1,
group_parameters = 0, joint_parameters = 18)
### Option 2 for the a-path: change in trend between the two phases
### (beta.3M)
results.trend <- bain(estimate,
                      "beta.3M > 0 & beta.4Y >0;
                      beta.3M < 0;
                      beta.4Y < 0",
                      n = N, Sigma = cov1,
                      group_parameters = 0, joint_parameters = 18)

###### Return output of Bain analysis
results.level
results.trend

##### Function to recompute posterior probabilities to include the
##### complement
complement.probs <- function(x, comp.hyps = NULL){
  # comp.hyps is a vector containing the ID number of the
  # hypothsis/es of which the complement should be included
  # in the posterior probabilities
  if(class(x) != "bain") stop("please provide a Bain
                              output object as x")
  if(is.null(comp.hyps)) return(x$fit)
```

```
  oldnames <- row.names(x$fit)
  for(i in comp.hyps){
    x$fit <- rbind(x$fit, c(1-x$fit[i, 1:6],"BF" = 1/x$fit[i, 7],
"PMPa" = NA, "PMPb" = NA))
  }
  row.names(x$fit) = c(oldnames, paste0("Hc", comp.hyps))
  PMPc.unst <- x$fit$Fit / x$fit$Com
  PMPc <- PMPc.unst/sum(PMPc.unst, na.rm = TRUE)
  x$fit <- cbind(x$fit, "PMPc" = PMPc)
  return(x$fit)
}

##### Use function to compute posterior probabilities for all
##### hypotheses including the complement of H1 (comp.hyps = 1)
postprob.level <- complement.probs(results.level, comp.hyps = 1)
postprob.trend <- complement.probs(results.trend, comp.hyps = 1)
postprob.level
postprob.trend
```

References

Adcock, C. J. (1997). Sample size determination: A review. *Journal of the Royal Statistical Society*, Series D, 46, 261–283.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csàki (Eds.), *Proc. 2nd Int. Symp. Information Theory* (pp. 267–281). Budapest: Akademiai kiado.

Angrilli, A., Cherubini, P., Pavese, A., & Manfredini, S. (1997). The influence of affective factors on time perception. *Perception & Psychophysics*, 59(6), 972–982. Retrieved from https://doi.org/10.3758/bf03205512

Babik, I., Cunha, A. B., Moeyaert, M., Hall, M. L., Paul, D. A., Mackley, A., & Lobo, M. A. (2019). Feasibility and effectiveness of intervention with the playskin lift exoskeletal garment for infants at risk. *Physical Therapy*, 99(6), 666–676. Retrieved from https://doi.org/10.1093/ptj/pzz035

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.

Berger, J. O., Boukai, B., & Wang, J. (1997). Unified frequentist and bayesian testing of a precise hypotheses. *Statistical Science*, 12(3), 133–148.

Béland, S., Klugkist, I., Raîche, G., & 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(2), 122–126.

Bruder, G., & Steinicke, F. (2014). Time perception during walking in virtual environments. In *Proceedings of ieee virtual reality* (pp. 67–68).

Burle, B., & Casini, L. (2001). Dissociation between activation and attention effects in time estimation: Implications for internal clock models. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 195–205. Retrieved from https://doi.org/10.1037/0096-1523.27.1.195

Cattell, R. B. (1952). The three basic factor-analytic research designs—their interrelations and derivatives. *Psychological Bulletin*, 49(5), 499–520.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, 19(4), 387–400. Retrieved from https://doi.org/10.1177/002246698501900404

Chirico, A., D'Aiuto, M., Pinto, M., Milanese, C., Napoli, A., Avino, F., ... Lucidi, F. (2016). Intelligent interactive multimedia systems and services. In G. Pietro, L. Gallo, R. Howlett, & L. Jain (Eds.) (pp. 731–738). Smart Innovation, Systems; Technologies.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum.

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

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

Craig, C. C. (1936). On the frequency function of xy. *The Annals of Mathematical Statistics*, 7(1), 1–15. Retrieved from https://doi.org/10.1214/aoms/1177732541

Cuperus, A. A., & Ham, I. J. van der. (2016). Virtual reality replays of sports performance: Effects on memory, feeling of competence, and performance. *Learning and Motivation*, 56, 48–52. Retrieved from https://doi.org/10.1016/j.lmot.2016.09.005

Cuperus, A. A., Keizer, A., Evers, A. W., Houten, M. M. van den, Teijink, J. A., & Ham, I. J. van der. (2018). Manipulating spatial distance in virtual reality: Effects on treadmill walking performance in patients with intermittent claudication. *Computers in Human Behavior*, 79, 211–216. Retrieved from https://doi.org/10.1016/j.chb.2017.10.037

Darnieder, W. F. (2011). Bayesian methods for data-dependent priors. Ohio State University, Columbus, OH.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypotheses testing. *Journal of Statistical Planning and Inference*, 124, 121–144. Retrieved from https://doi.org/10.1007/s11749-006-0017-7

De Santis, F. (2007). Alternative Bayes factors: Sample size determination and discriminatory power assessment. *Test*, 16, 504–522. Retrieved from https://doi.org/10.1016/S0378-3758(03)00198-8

Droit-Volet, S., & Meck, W. H. (2007). How emotions colour our perception of time. *Trends in Cognitive Sciences*, 11(12), 504–513. Retrieved from https://doi.org/10.1016/j.tics.2007.09.008

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, 16(1), 143–149. Retrieved from https://doi.org/10.3758/bf03203267

Felnhoger, K., A. (2015). Is virtual reality emotionally arousing? Investigation five emotion inducing virtual park scenarios. *International Journal of Human-Computer Studies*, 82, 48–56.

Finnegan, O., D. (2016). Compensating for distance compression in audiovisual virtual environments using incongruence. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 200–212). Association for Computing Machinery.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470). Retrieved from https://doi.org/10.1198/016214505000000105

Gaynor, S. T., & Harris, A. (2008). Single-participant assessment of treatment mediators. *Behavior Modification*, 32(3), 372–402. Retrieved from https://doi.org/10.1177/0145445507309028

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), 1360–1383. Retrieved from https://doi.org/10.1214/08-AOAS 191

Gu, X., Hoijtink, H., Mulder, J., & Lissa, C. van. (2019). *Bain: Bayes factors for informative hypotheses* (R package version 0.2.0). R package version 0.2.0.

Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19, 511–527. Retrieved from https://doi.org/10.1037/met0000017

Gu, X., Mulder, J., & Hoijtink, H. (2017). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71(2), 229–261. Retrieved from https://doi.org/10.1111/bmsp.12110

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, 22(4), 779–798. Retrieved from https://doi.org/10.1037/met0000156

Hamaker, L., E. (2012). Handbook of research methods for studying daily life. In M. R. Mehl & S. Conner (Eds.) (pp. 43–61). New York, NY: Guilford.

Ham, K. van der, I. J. M. (2019). Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence. *Computers in Human Behavior*, 94, 77–81. Retrieved from https://doi.org/https://doi.org/10.1016/j.chb.2019.01.005

Hancock, P., & Rausch, R. (2010). The effects of sex, age, and interval duration on the perception of time. *Acta Psychologica*, 133(2), 170–179. Retrieved from https://doi.org/10.1016/j.actpsy.2009.11.005

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, 3(3), 224–239. Retrieved from https://doi.org/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, 4(4), 324–341. Retrieved from https://doi.org/10.1002/jrsm.1086

Hofstee, W. K. B. (1984). Methodological decision rules as research policies: A betting reconstruction of empirical research. *Acta Psychologica*, 56, 93–109.

Hoijtink, H. (2012). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists*. Boca Raton: Chapman & Hall/CRC.

Hoijtink, H., Gu, X., & Mulder, J. (n.d.). *Bain, multiple group bayesian evaluation of informative hypotheses.* Retrieved from https://informative-hypotheses.sites.uu.nl/wp-content/uploads/sites/23/2018/01/MGBain.pdf

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian Evaluation of Informative Hypotheses*. New York: Springer.

Hoijtink, H., Mulder, J., Lissa, C. van, & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*. Retrieved from https://doi.org/10.1037/met0000201

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71(2), 165–179. Retrieved from https://doi.org/10.1177/001440290507100203

Hout, M. van den, Gangemi, A., Mancini, F., Engelhard, I. M., Rijkeboer, M. M., Dams, M. van, & Klugkist, I. (2014). Behavior as information about threat in anxiety disorders: A comparison of partients with anxiety disorders and non-anxious controls. *Journal of Behavior Therapy and Experimental Psychiatry*, 45, 489–495. Retrieved from https://doi.org/10.1016/j.jbtep.2014.07.002

Indovina, P., Barone, D., Gallo, L., Chirico, A., Pietro, G. D., & Antonio, G. (2018). Virtual reality as a distraction intervention to relieve pain and distress during medical procedures. *The Clinical Journal of Pain*, 1. Retrieved from https://doi.org/10.1097/ajp.0000000000000599

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 696–701. Retrieved from https://doi.org/10.1371/journal.pmed.0020124

Jarosz, A. F., & Wiley, J. (2017). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1), Article 2. Retrieved from https://doi.org/10.7771/1932-6246.1167

JASP Team. (2018). JASP Version 0.19.0.0)[Computer software]. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1998). *Theory of probability* (3rd ed.). Oxford University Press.

Johnson, S. R., Tomlinson, G. A., Hawker, H. A., Granton, J. T., & Feldman, B. M. (2010). Methods to elicit belief for Bayesian priors: A systematic review. *Journal of Clinical Epidemiology*, 63, 355–369. Retrieved from https://doi.org/10.1016/j.jclinepi.2009.06.003

Judd, C. M., & Kenny, D. A. (1981). Process analysis. *Evaluation Review*, 5(5), 602–619. Retrieved from https://doi.org/10.1177/0193841x8100500502

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

Kazdin, A. E. (2011). *Single-case research designs* (Second Edition). New York, NY: Oxford University Press.

Klaassen, F. (2019). *BayesianPower: Sample size and power for comparing inequality constrained hypotheses* (R package version 0.1.6).

Klaassen, F., Hoijtink, H., & Gu, X. (n.d.). *The power of informative hypotheses*. Retrieved from https://doi.org/10.31219/osf.io/d5kf3

Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H., & Hoijtink, H. (2017). All for one or some for all? Evaluating informative hypotheses using multiple N =1 studies. *Behavior Research Methods*, 50(6), 2276–2291. Retrieved from https://doi.org/10.3758/s13428-017-0992-5

Klimecki, O. M., Mayer, S. V., Jusyte, A., Scheeff, J., & Schönenberg, M. (2016). Empathy promotes altruistic behavior in economic interactions. *Scientific Reports*, 6(31961), 1–5. Retrieved from https://doi.org/10.1038/srep31961

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10(4), 477–493. Retrieved from https://doi.org/10.1037/1082-989X.10.4.477

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables. *Psychological Methods*, 15(3), 281–299. Retrieved from https://doi.org/10.1037/a0020137

Klugkist, I., Post, L., Haarhuis, F., & Wesel, F. van. (2014). Confirmatory methods, or huge samples, are required to obtain power for the evaluation of theories. *Open Journal for Statistics*, 4, 710–725. Retrieved from https://doi.org/10.4236/ojs.2014.49066

Klugkist, I., Wesel, F. van, & Bullens, J. (2011). Do we know what we test and to we test what we want to know? *International Journal of Behavioral Development*, 35(6), 550–560. Retrieved from https://doi.org/10.1177/0165025411425873

Kluytmans, A., Van de Schoot, R., Zedelius, C., Veling, H., Aarts, H., & Hoijtink, H. (n.d.). *Bayesian sequential evaluation of simple order constraints using dichotomous within-subject data*. Retrieved from Unpublished manuscript

Knapp, J. M., & Loomis, J. M. (2004). Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoperators and Virtual Environments*, 13(5), 572–577. Retrieved from https://doi.org/10.1162/1054746042545238

Konijn, E. A., Van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, 9(4), 280–302. Retrieved from https://doi.org/10.1080/19312458.2015.1096332

Kopp, B., Seer, C., Lange, F., Kluytmans, A., Kolossa, A., Fingscheidt, T., & Hoijtink, H. (2016). P300 amplitude variations, prior probabilities, and llikelihood: A Bayesian ERP study. *Cognitive, Affective and Behavioral Neuroscience*, 16, 911–928. Retrieved from https://doi.org/10.3758/s13415-016-0442-3

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and planning from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. Retrieved from https://doi.org/10.2139/ssrn.2606016

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

Lang, B., P. J. (1997). International affective picture system (IAPS): Technical manual and affective ratings.

Linkenauger, S. A., Bülthoff, H. H., & Mohler, B. J. (2015). Virtual arms reach influences perceived distances but only after experience reaching. *Neuropsychologia*, 70, 393–401. Retrieved from https://doi.org/10.1016/j.neuropsychologia.2014.10.034

Lomnicki, Z. A. (1967). On the distribution of products of random variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 29(3), 513–524. Retrieved from https://doi.org/10.1111/j.2517-6161.1967.tb00713.x

Maanen, L. van, Forstmann, B. U., Keuken, M. C., Wagenmakers, E. J., & Heathcote, A. (2016). The impact of MRI scanner environment on perceptual decision-making. *Behavior Research Methods*, 48, 184–200. Retrieved from https://doi.org/10.3758/s13428-015-0563-6

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86–92. Retrieved from https://doi.org/10.1027/1614-2241.1.3.86

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Routledge.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83–104. Retrieved from https://doi.org/10.1037/1082-989x.7.1.83

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, 39(1), 99–128. Retrieved from https://doi.org/10.1207/s15327906mbr3901_4

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, 30(1), 41–62. Retrieved from https://doi.org/10.1207/s15327906mbr3001_3

Manolov, R., & Moeyaert, M. (2016). How can single-case data be analyzed? Software resources, tutorial, and reflections on analysis. *Behavior Modification*, 41(2), 179–228. Retrieved from https://doi.org/10.1177/0145445516664307

Maric, M., Heyne, D. A., MacKinnon, D. P., Widenfelt, B. M. van, & Westenberg, P. M. (2012). Cognitive mediation of cognitive-behavioural therapy outcomes for anxiety-based school refusal. *Behavioural and Cognitive Psychotherapy*, 41(5), 549–564. Retrieved from https://doi.org/10.1017/s1352465812000756

Maric, M., Prins, P.J.M., & Ollendick, T. (Eds.). (2015). *Moderators and mediators of youth treatment outcomes*. New York: Oxford University Press.

Matthews, W. J., & Meck, W. H. (2016). Temporal cognition: Connecting subjective time to perception, attention, and memory. *Psychological Bulletin*, 142(8), 865–907. Retrieved from https://doi.org/10.1037/bul0000045

Merkle, E. C., & Rosseel, Y. (2018). Blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85(4). Retrieved from https://doi.org/10.18637/jss.v085.i04

Miočević, M., MacKinnon, D. P., & Levy, R. (2017). Power in bayesian mediation analysis for small sample research. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(5), 666–683. Retrieved from https://doi.org/10.1080/10705511.2017.1312407

Miočević, &. V. de S., M. (2019). Advanced research methods and statistics for the behavioral and social sciences. In J. E. & A. L. Nichols (Ed.). Cambridge, UK: Cambridge University Press.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Noortgate, W. V. den. (2013a). The three-level synthesis of standardized single-subject experimental data: A monte carlo simulation study. *Multivariate Behavioral Research*, 48(5), 719–748. Retrieved from https://doi.org/10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Noortgate, W. V. den. (2013b). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, 82(1), 1–21. Retrieved from https://doi.org/10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Noortgate, W. V. den. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-subject experimental design research. *Behavior Modification*, 38(5), 665–704. Retrieved from https://doi.org/10.1177/0145445514535243

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement Interdisciplinary Research and Perspectives*, 2(4), 201–218. Retrieved from https://doi.org/10.1207/s15366359mea0204\textunderscore 1

Monin, B., Sawyer, P. J., & Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, 95(1), 76–93. Retrieved from https://doi.org/10.1037/0022-3514.95.1.76

Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal of Experimental Social Psychology*, 18, 395–415. Retrieved from https://doi.org/10.1016/0022-1031(82)90062-2

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6–18. Retrieved from https://doi.org/10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (R package version 0.9.12-4.2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Mulder, J. (2014). Bayes factors for testing inequality constrained hypotheses: Issues with prior specification. *British Journal of Mathematical and Statistical Psychology*, 67, 153–171. Retrieved from https://doi.org/10.1111/bmsp.12013

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. Retrieved from https://doi.org/10.1016/j.jspi.2009.09.022

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

Mulder, J., Klugkist, I., Schoot, R. van de, Meeuw, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, 53, 530–546. Retrieved from https://doi.org/10.1016/j.jmp.2009.09.003

Mulder, J., & Wagenmakers, E. J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, 72, 1–5. Retrieved from https://doi.org/10.1016/j.jmp.2016.01.002

Neuman, W. L. (2011). *Social research methods: Qualitative and quantitative approaches* (7th ed.). Boston: Allyn; Bacon.

Ng, M. Y., & Weisz, J. R. (2015). Annual research review: Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry*, 57(3), 216–236. Retrieved from https://doi.org/10.1111/jcpp.12470

Noulhiane, M., Mella, N., Samson, S., Ragot, R., & Pouthas, V. (2007). How emotional auditory stimuli modulate time perception. *Emotion*, 7(4), 697–704. Retrieved from https://doi.org/10.1037/1528-3542.7.4.697

O'Hagan, A., Buck,.. C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., ... Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons, Ltd. Retrieved from https://doi.org/10.1002/0470033312

O'Hagan, A., & Perichhi, L. (2012). Bayesian heavy-tailed models and conflict resolution: A review. *Brazilian Journal of Probability and Statistics*, 26(4), 372–401. Retrieved from https://doi.org/10.1214/11-BJPS164

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, 40(4), 357–367. Retrieved from https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, 75(2), 135–150. Retrieved from https://doi.org/10.1177/001440290907500201

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-u. *Behavior Therapy*, 42(2), 284–299. Retrieved from https://doi.org/10.1016/j.beth.2010.08.006

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530. Retrieved from https://doi.org/10.1177/1745691612465253

Plummer, B., M. (2018). Coda: Output analysis and diagnostics for mcmc. R package version 0.19-2.

Plummer, M. (2003). *JAGS: A program for analysis of bayesian graphical models using gibbs sampling.*

Plummer, M. (2018). Rjags: Bayesian graphical models using mcmc. R package version 4-8.

Rac-Lubashevsky, R., & Kessler, Y. (2016). Decomposing the n-back task: An individual differences study using the reference-back paradigm. *Neuropsychologia*, 90, 190–199. Retrieved from https://doi.org/10.1016/j.neuropsychologia.2016.07.013

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from http://www.R-project.org/

Reyes, E. M., & Ghosh, S. K. (2013). Bayesian average error-based approach to sample size calculations for hypotheses testing. *Journal of Biopharmaceutical Statistics*, 23, 569–588. Retrieved from https://doi.org/10.1080/10543406.2012.755994

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641. Retrieved from https://doi.org/10.1037/0033-2909.86.3.638

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10), 1276–1284. Retrieved from https://doi.org/10.1037/0003-066X.44.10.1276

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301–308. Retrieved from https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., & Wagenmakers, E. J. (2016). The interplay between subjectivity, statstical practice, and psychological science. *Collabra*, 2(1), 1–12. Retrieved from https://doi.org/10.1525/collabra.28

Schatzschneider, C., Bruder, G., & Steinicke, F. (2016). Who turned the clock? Effects of manipulated zeitgebers, cognitive load and immersion on time estimation. *IEEE Transactions on Visualization and Computer Graphics*, 22(4), 1387–1395. Retrieved from https://doi.org/10.1109/tvcg.2016.2518137

Scherbaum, C. A., & Ferreter, J. M. (2008). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, 12(2), 347–367. Retrieved from https://doi.org/10.1177/1094428107308906

Schneider, S. M., Kisby, C. K., & Flint, E. P. (2010). Effect of virtual reality on time perception in patients receiving chemotherapy. *Supportive Care in Cancer*, 19(4), 555–564. Retrieved from https://doi.org/10.1007/s00520-010-0852-7

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25, 128–142. Retrieved from https://doi.org/10.3758/s13423-017-1230-7

Scott, J. G., & Berger, J. O. (2010). Bayers and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5), 2587–2619. Retrieved from https://doi.org/10.1214/10-AOS792

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. *Remedial and Special Education*, 8(2), 24–33. Retrieved from https://doi.org/10.1177/074193258700800206

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, 52(2), 123–147. Retrieved from https://doi.org/10.1016/j.jsp.2013.11.005

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 2(3), 188–196. Retrieved from https://doi.org/10.1080/17489530802581603

Sharar, S. R., Alamdari, A., Hoffer, C., Hoffman, H. G., Jensen, M. P., & Patterson, D. R. (2016). Circumplex model of affect: A measure of pleasure and arousal during virtual reality distraction analgesia. *Games for Health Journal*, 5(3), 197–202. Retrieved from https://doi.org/10.1089/g4h.2015.0046

Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality and shape constraints.* London: Wiley.

Sinharay, S. (2004). Experiences with markov chain monte carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, 29(4), 461–488. Retrieved from https://doi.org/10.3102/10769986029004461

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17(4), 510–550. Retrieved from https://doi.org/10.1037/a0029312

Stefanucci, J. K., Creem-Regehr, S. H., Thompson, W. B., Lessard, D. A., & Geuss, M. N. (2015). Evaluating the accuracy of size perception on screen-based displays: Displayed objects appear smaller than real objects. *Journal of Experimental Psychology: Applied*, 21(3), 215–223. Retrieved from https://doi.org/10.1037/xap0000051

Stephan, K. E., & Penny, W. D. (2007). Dynamic Causal Models and Bayesian Selection. In K. Friston, J. Ashbumer, S. Kievel, T. Nichols, & W. Penny (Eds.), *Statistical Parametric Mapping: The Analysis of Functional Brain Images* (pp. 577–585). Academic Press.

Tarlow, K. R. (2016). An improved rank correlation effect size statistic for single-case designs: Baseline corrected tau. *Behavior Modification*, 41(4), 427–467. Retrieved from https://doi.org/10.1177/0145445516676750

Thompson, B. (2004). The "significance" crisis in psychology and education. *The Journal of Socio-Economics*, 33, 607–613. Retrieved from https://doi.org/10.1016/j.socec.2004.09.034

Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests tdoes not suffice: A plea for plausibility. *Psychonomic Bulletin & Review*, 25, 548–559. Retrieved from https://doi.org/10.3758/s13423-018-1447-4

Vanbrabant, L., & Rosseel, Y. (2020). Small sample size solutions: A guide for applied researchers and practitioners. In Van de Schoot R. & M. Miočević (Eds.). Routledge.

Vanbrabant, L., Schoot, R. van de, & Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for anova and regression. *Frontiers in Psychology*, 5. Retrieved from https://doi.org/10.3389/fpsyg.2014.01565

VandenBos, G. R. (Ed.). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington DC, USA: American Psychological Association.

van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, 18(3), 325–346. Retrieved from https://doi.org/10.1521/scpq.18.3.325.22577

van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, 35(1), 1–10. Retrieved from https://doi.org/10.3758/bf03195492

van den Noortgate, W., & Onghena, P. (2007). The aggregations of single-case research using hierarchical linear models. *The Behavior Analist Today*, 8(2), 196–209.

van de Schoot, R., Hoijtink, H., & Romeijn, J. W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, 2. Retrieved from https://doi.org/10.3389/fpsyg.2011.00024

Van Erp, S. (2020). Small sample size solutions: A guide for applied researchers and practitioners. In Van de Schoot R. & M. Miočević (Eds.). Routledge.

Vannest, K. J., Peltier, C., & Haas, A. (2018). Results reporting in single case experiments and single case meta-analysis. *Research in Developmental Disabilities*, 79, 10–18. Retrieved from https://doi.org/10.1016/j.ridd.2018.04.029

Villa, C., & Walker, S. (2015). An objective Bayesian criterion to determine model prior probabilities. *Scandinavian Journal of Statistics*, 42, 947–966. Retrieved from https://doi.org/10.1111/sjos.12145

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Maas, H. L. J. van der. (2011). Why psychologists must change the way they analyze their data: The case of Psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432. Retrieved from https://doi.org/10.1037/a0022790

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, 46, 185–191.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. J. (2012). Testing bayesian hypothesis test for anova designs. *The American Statistician*, 66(2), 104–111. Retrieved from https://doi.org/10.1080/00031305.2012.695956

Wiessenekker, M. (2019). Efficacy of child and adolescent therapies: A meta-analysis of single-case studies.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds or testig a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*. Retrieved from https://doi.org/10.1177/2515245918767122

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2008). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, 44(1), 18–28. Retrieved from https://doi.org/10.1177/0022466908328009

Woodcock, J. (2007). The prospects for "personalized medicine" in drug development and drug therapy. *Clinical Pharmacology and Therapeutics*, 81(2), 164–169. Retrieved from https://doi.org/10.1038/sj.clpt.6100063

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, 14(4), 301–322. Retrieved from https://doi.org/10.1037/a0016972

Zedelius, C. M., Veling, H., & Aarts, H. (2011). Boosting or choking – How conscious and unconscious reward processing modulate the active maintenane of goal-relevant information. *Consciousness and Cognition*, 20, 355–362. Retrieved from https://doi.org/10.1016/j.concog.2010.05.001

Wetenschappelijke samenvatting

Binnen de sociale en gedragswetenschappen neemt het gebruik van Bayesiaans informatief hypothese toetsen toe (Mulder & Wagenmakers, 2016). Informatieve hypotheses beschrijven gerichte verwachtingen van een onderzoeker die op basis van theorie worden gespecificeerd (e.g. Hoijtink, 2012; Klugkist et al., 2005). Door deze specifieke hypotheses te evalueren in plaats van standaard nul en alternatieve hypotheses kan de onderzoeksvraag beter worden beantwoord. Bayesiaanse statistiek kan gebruikt worden om het relatieve bewijs voor meerdere hypotheses te kwantificeren door middel van Bayes factors. Het updaten van kennis centraal in de Bayesiaanse statistiek. De combinatie van informatieve hypotheses en Bayesiaans updaten is nauw verbonden met de onderzoekscyclus. Deze beschrijft hoe uit theorie verwachtingen worden geformuleerd, die vervolgens aan de hand van data kunnen worden getoetst. De conclusies van statistische analysese kunnen dan weer nieuwe theoriën inspireren of andere richtingen aan de verwachting geven. Met de ontwikkeling van toegangelijke software en de uitbreiding van bestaande statistische software groeit het gebruik van informatief Bayesiaans hypothese toetsen.

Het toegenomen gebruik van Bayesiaans informatief hypothese toetsen resulteert in praktische, filosofische en methodologische vraagstukken. De Bayes factor heeft een duidelijke interpretatie als het relatieve bewijs voor twee hypotheses. Minder eenduidig is welke conclusies wel en niet op basis van een Bayes factor getrokken kunnen worden. Hoe sterk moet het relatieve bewijs zijn voordat het overtuigend genoeg is? Moet de sterkte van de conclusie afhangen van de hoeveelheid data die is gebruikt om tot het bewijs te komen? Andere vragen gaan over het gebruik en het opstellen van informatieve hypotheses. Elke verwachting kan in een hypotheses worden gevat. Zodoende is er een groot aantal hypotheses dat potentieel interessant kan zijn. Hoe kan je als onderzoeker de keuze voor bepaalde hypotheses maken, en andere buiten beschouwing laten? Om informatieve hypothesen Bayesiaans te toetsen moet een onderzoeker beslissingen nemen over onder andere de hypotheses, de prior verdelingen, en het interpreteren van Bayes factors. Voor veel van deze beslissingen zijn geen richtlijnen beschikbaar. In deze dissertatie worden een aantal van deze vraagstukken behandeld.

In Hoofdstuk 2 wordt het verband tussen steekproefgrootte en conditionele en onconditionele foutkansen besproken in de context van Bayesiaans informatieve hypotheses toetsen. Conditionele foutkansen, de kans op een foute beslissing gegeven de huidige data, zijn van de Bayes factor af te leiden. Deze foutkansen hebben vaak de focus in Bayesiaans hypothese toetsen. Onconditionele foutkansen bestaan ook voor Bayes factors. Dit is de kans op een foute beslissing vóórdat data zijn geobserveerd. Beide kansen hangen samen met de steekproefgrootte binnen een onderzoek. Het bepalen van de

steekproefgrootte is een belangrijke stap bij het opzetten van een onderzoek. Hoofdstuk 2 beschrijft verschillende manieren waarop de steekproefgrootte bepaald kan worden als een onderzoeker informatieve hypotheses Bayesiaans wil toetsen. Zo kan rekening gehouden worden met een beoogde sterkte van bewijs, of met een maximaal toelaatbare (on)conditionele foutkans. Dit hoofdstuk laat zien dat er veel keuzes zijn die gemaakt moeten worden door de onderzoeker, en geeft handvatten om deze keuzes te maken.

Hoofdstuk 3 en 4 bespreken het niveau van de hypothese en de onderzoeksvraag. Standaard worden de meeste hypothese toetsen op het groepsniveau uitgevoerd: is er gemiddeld genomen in de populatie een verschil tussen mannen en vrouwen, of werkt een medicijn gemiddeld genomen beter dan de placebo. Echter, in veel gevallen is de interesse van de onderzoeker op het niveau van het individu: zou het deze patient meer helpen om een medicijn te krijgen of een placebo? Is deze leerling gebaat bij bijles? Hoofdstuk 3 presenteert een methode om te evalueren of een hypothese niet *gemiddeld*, maar *voor iedereen* geldt. Hiervoor worden individuele datasets onafhankelijk geanalyseerd, en deze informatie kan met behulp van Bayesiaanse statistiek worden samengenomen. Hoofdstuk 4 beschrijft stap voor stap hoe een onderzoeker deze analyse ook zelf zou kunnen uitvoeren. Deze vorm van informatie-synthese over meerdere individuele hypothese toetsen geeft antwoord op een specifieke vraag, namelijk: wat is de beste hypothese voor iedereen? Opnieuw laat dit zien hoe de keuzes van een onderzoeker de mogelijke conclusies van een onderzoek beinvloeden.

In Hoofdstuk 5 wordt het filosofische aspect van het voortdurend updaten van kennis verder onder de loep genomen. Bayesiaanse statistiek word vaak gepromoot vanwege de mogelijkheid voortdurend voort te bouwen op bestaande kennis. In de praktijk gebeurt dit nog weinig, omdat het vaak moeilijk is om de eerdere kennis betekenisvol te kwantificeren. Als een gevolg blijven veel conclusies op het niveau van het beschrijven van het bewijs in de huidige data, zonder daadwerklijk de kennis te updaten. In Hoofdstuk 5 word het concept van de prior kans van een hypothese ontleed. Aan de hand van deze definitie is een procedure ontwikkeld, gepresenteerd en getest, om hier betekenisvol een getal aan toe te kennen.

Hoofdstuk 6 beschrijft de beschikbare software die is ontwikkeld, om de methoden uit Hoofdstuk 2, 3 en 4 ook zelf toe te kunnen passen. Hoofdstuk 7 en 8 beschrijven onderzoeken waar daadwerkelijk Bayesiaans informatief hypothese toetsen is toegepast. Deze hoofdstukken illustreren de diversiteit en flexibiliteit aan conclusies die deze methode met zich meebrengt. Zo is in Hoofdstuk 7 een repeated measures model gebruikt, en worden informatieve hypotheses op het groepsniveau getoetst. In Hoofdstuk 8 wordt aan de hand van een individueel mediatie model bekeken of er voor 1 persoon een gemedieerd effect kan worden waargenomen. Dit zou uitgebreid kunnen worden met de methoden uit Hoofdstuk 3 om te kijken of hetzelfde effect voor alle proefpersonen gevonden kan worden.

Bayesiaans informatief hypothese toetsen biedt veel mogelijkheden aan de sociale en gedragswetenschappen. Zo kunnen specifieke verwachtingen getoetst worden en kan kennis over theorien blijven groeien naarmate meer data verzameld wordt. Deze mogelijkheden komen ook met veel verantwoordelijkheden en keuzes. Het is van belang dat de keuzes die gemaakt worden in het proces, worden toegelicht en onderbouwd. Daarnaast valt er ook nog veel te leren over het belang van de keuze voor bepaalde hypotheses, steekproefgrootte of het niveau van de analyse. Zolang we onze kennis ook hierover blijven updaten en laten groeien, zal ook het gebruik van Bayesiaans informatief hypothese toetsen toenemen.

About the author

Fayette Klaassen was born on May 13th 1992 in Amsterdam, the Netherlands. She obtained her BA cum laude in liberal arts and sciences from Amsterdam University College in 2013, with a major in psychology. In her bachelor thesis she investigated the connections between research, policy and practice in Dutch education. She started with two research masters, in educational sciences and in applied methodology and statistics. Her interest in research methodology and statistics surmounted that for educational sciences. In 2015 she obtained her MSc cum laude at Utrecht University, with a master thesis on Bayesian hypothesis testing.

She continued after her master thesis with a PhD at Utrecht University in 2015. She participated in the Graduate Think Tank of Utrecht Univeristy part between 2015 and 2017, and was a member of the organizing committee of the 2016 graduate research conference 'Dare to cross-over', resulting from this Think Tank. Between 2016 and 2018 She was active as a PhD representative at the department of Methodology and Statistics, in the PhD council of the faculty of Social Sciences of Utrecht University and at the interuniversity graduate school of psychometrics and sociometrics (IOPS). She was selected to attend the LERU Graduate Summer School on behalf of Utrecht University in 2018. In 2019 she visited prof.dr. Jeffrey Rouder at University of California, Irvine, USA as part of her PhD program.

Publications

Klaassen, F. (in press). Combining evidence over multiple individual analyses. In R. Van de Schoot & M. Miocevic (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners.* Routledge.

Kuiling, J. M. E., **Klaassen, F.** & Hagenaars, M. A. (2019). The role of tonic immobility and control in the development of intrusive memories after experimental trauma. *Memory.* (Online ahead of print). doi:10.1080/09658211.2018.1564331

van der Ham, I. J. M., **Klaassen, F.**, van Schie, K., & Cuperus, A. (2019). Elapsed time estimates in virtual reality and the physical world: The role of arousal and emotional valence. *Computers in Human Behavior, 94*, pp.77-81. doi:10.1016/j.chb.2019.01.005

Klaassen, F., Zedelius, C., Veling, H., Aarts, H., & Hoijtink, H. (2017). All for One or Some for All? Evaluating Informative Hypotheses using Multiple $N = 1$ Studies. *Behavior Research Methods* doi:10.3758/s13428-017-0992-5

Hagenaars, M. A., Holmes, E. A., **Klaassen, F.**, & Elzinga, B. (2017). Tetris and word game affect intrusive memories when applied after memory reactivation 4 days post analogue trauma. *European Journal of Psychotraumatology. 8* (Sup 1).

Mulder, H. de, Hakemulder, F., van den Berghe, R., **Klaassen, F.** & van Berkum, J. (2017). Effects of exposure to literary narrative fiction: From book smart to street smart? *Scientific Study of Literature, 7* (1), pp. 129-169.

Cuperus, A. A., **Klaassen, F.**, Hagenaars, M. A., & Engelhard, I. M. (2017). A virtual reality paradigm as an analogue to real-life trauma: Its effectiveness compared with the trauma film paradigm. *European Journal of Psychotraumatology. 8* (Sup 1). doi:10.1080/20008198.2017.1338106

Dankwoord

Vier jaar geleden had ik geen flauw idee over wat me te wachten stond, en was ik onzeker over wat ik toe kon voegen. Zo anders voelt dat nu, aan het einde van de rit. Met trots en tevredenheid kijk ik naar wat ik heb geleerd en hoe ik ben gegroeid. In mijn proefschrift, maar ook persoonlijk. Met liefde en dankbaarheid kijk ik naar hoe ik ben gesteund, geholpen en gedragen tijdens frustraties, zoektochten en het verdriet van de afgelopen jaren. Zonder jullie: familie, vrienden en collega's, zou ik hier nu niet staan.

Dankjewel.

Mijn ouders, Joke en Henk-Willem. Jullie gaven me een liefde voor taal en onderzoeken, voor cijfers en creativiteit mee. Jullie moedigden mij aan mezelf uit te dagen en leerden me trots te zijn op mijn prestaties. Jullie gaven me de kracht om uberhaupt te beginnen.

Herbert. Dankjewel, voor alle tijd en aandacht die je voor mij had. Je hebt me de vrijheid gegeven om mijn eigen ideeën te volgen, mijn grenzen aan te geven en me te ontwikkelen. Je was altijd beschikbaar als ik wat wilde bespreken en je kritische blik en vragen – *Maar hoe weet je zeker dat het klopt?* – hielpen mij om dichter bij het probleem te komen. Voor mij was het een perfecte match waarbij ik zowel werd vrijgelaten en werd uitgedaagd.
Irene. Dankjewel dat ik altijd welkom was ideeën met je te delen, of mijn onderzoek met je te bespreken. Onze gesprekken over gedeelde interesses brengen altijd weer nieuwe energie en zin in mijn onderzoek en het belang ervan.

Anne and Oisín. You are without a doubt the two people I shared most of the ups and downs of PhD life with.
Anne, dankjewel voor je bodemloze drop-pot, onze fietstochtjes en uitstapjes. Voor je eeuwige stuiterende positieve energie om me weer op te laden als ik er even doorheen zit en de knuffels wanneer ik zelf nog niet weet dat ik die nodig heb.
Oisín, I'm forever grateful for the day you decided we should be friends. Thank you for always being there to share the miseries, solve problems and rant about frustrations.

Mijn dank gaat ook uit naar mijn co-auteurs. Deze samenwerkingen gaven me nieuwe energie en waardering voor mijn werk. Hannah, Henk, Muriël, Anne en Ineke bedankt voor onze fijne gesprekken over Bayes factors, hypotheses en de praktische problemen van echte data. Jeff, thank you for your hospitality during my time in Irvine, for the new perspectives and stimulating discussions about our research, together with Julia. Milica, thank you for our talks about and not about research.

Alle collega's van de M&S department, bedankt voor de fijne werksfeer. Kees en Jolien, voor het afwisselen van discussies over Bayes met de grote vragen in het leven: breekt

een ei als je deze met parachute uit het raam gooit? Flip, Irma, Sanne en Yana, voor het ontspannen tijdens office yoga. Sanne, voor ons gedeelde klim en zingplezier. Ayoub, Corine, Duco, Erik-Jan, Hidde, Karlijn, Lientje, Sjoerd en alle andere PhDs en judo's die aansloten bij de 12-uur lunch met Jenga competities, statistische raadsels en discussies over broodje versus boterham, dankjewel voor de gezelligheid en de aandacht.

Spark, Rianne, Lysanne en Imke. Wat was het fijn om samen met jullie ons schrijfproces onder de loep te nemen op de trap in de Village en te bedenken wat onze doelen en niet-doelen zijn.

Op zoveel momenten hebben jullie, mijn familie, vrienden en collega's, eraan bijgedragen dat ik mijn hoofd leeg kon maken, me kon ontspannen of lekker plezier kon hebben. Op maandagavond zingen met Birdpack, elke zesde zondag wandelen in Amsterdam, op de racefiets door de Utrechtse polder en heuvels, een etentje of spelletjescompetitie, kopjes thee en koffie, wijntjes en biertjes, naar theater, musea of ballet, de natuur in, op vakantie of gewoon even samen te zijn.

Coen, Emma, Francesca, Jelte, Merijn, Nina, Paula, Simon, Zoé, en de anderen van de Second Floor. Dankjewel voor hoe het nog altijd zo vertrouwd voelt als we samen afspreken. Voor de betrokkenheid en voor het bewijs dat tijd en afstand er niet toe doen.

Mijn vriendinnen uit Castricum. Amber, Eline, Femke, Jesminne, Jorien en Loes. Dankjewel voor het telkens weer vragen waar mijn onderzoek ook alweer over gaat, en dankjewel voor het dan weer vergeten, zodat we samen kunnen genieten van het leven.

Millitza, Oisín, Roline en Thomas. Dankjewel voor alle tripjes, spelletjesavonden, etentjes en uitjes. Voor hoe jullie altijd voor me klaarstaan, no matter what.

Sean en Yannick. Thank you for always being happy to discuss science, philosophy or statistics with me, en voor alles daaromheen.

Mirjam. Dankjewel voor je hulp in het maken van de omslag van dit proefschrift. Voor onze nichtjes-vriendschap en creatieve-knusheid.

Laura en Loes. Jullie grenzeloze trots, aanmoedigingen en liefde geven me kracht. Dankjewel dat ik bij jullie altijd thuis kan komen, en thuis ben.

Wij vijf. Mijn hart loopt over van liefde voor jullie.
Lentine en Mariëlle. Jullie zijn mijn zon en mijn maan. Dankjewel voor alles wat we delen, waar we in verschillen of een spiegel zijn voor elkaar.
Lieve papa. Je bent mijn grootste supporter en uitdager. Van advies tot avontuur, dankjewel dat je er altijd voor me bent.
Lieve mama. Je bent niet meer hier, maar toch altijd bij me. In elke volgende stap en in elke nieuwe dag.

♥