Behav Res (2018) 50:2276–2291
https://doi.org/10.3758/s13428-017-0992-5

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

Fayette Klaassen[1] · Claire M. Zedelius[2] · Harm Veling[3] · Henk Aarts[4] · Herbert Hoijtink[1,5]

Published online: 15 December 2017

Abstract Analyses are mostly executed at the population level, whereas in many applications the interest is on the individual level instead of the population level. In this paper, multiple $N = 1$ experiments are considered, where participants perform multiple trials with a dichotomous outcome in various conditions. Expectations with respect to the performance of participants can be translated into so-called informative hypotheses. These hypotheses can be evaluated for each participant separately using Bayes factors. A Bayes factor expresses the relative evidence for two hypotheses based on the data of one individual. This paper proposes to "average" these individual Bayes factors in the gP-BF, the average relative evidence. The gP-BF can be used to determine whether one hypothesis is preferred over another for all individuals under investigation. This measure provides insight into whether the relative preference of a hypothesis from a pre-defined set is homogeneous over individuals. Two additional measures are proposed to support the interpretation of the gP-BF: the evidence rate (ER), the proportion of individual Bayes factors that support the same hypothesis as the gP-BF, and the stability rate (SR), the proportion of individual Bayes factors that express a stronger support than the gP-BF. These three statistics can be used to determine the relative support in the data for the informative hypotheses entertained. Software is available that can be used to execute the approach proposed in this paper and to determine the sensitivity of the outcomes with respect to the number of participants and within condition replications.

Keywords Bayes factor · Informative hypotheses · $N = 1$ studies · Within-subject experiment

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 (see Kluytmans et al., 2014, for a pilot study into using informative hypothesis in the context of multiple $N = 1$ studies). In a within-subject experiment each person $i = 1, ..., P$ is exposed to the same set of experimental conditions $j = 1, \ldots, J$. By conducting R

✉ Fayette Klaassen
klaassen.fayette@gmail.com

[1] Department of Methodology and Statistics, Utrecht University, PO Box 80140, 3508 TC Utrecht, The Netherlands

[2] Department of Psychology, University of California, Santa Barbara, CA, USA

[3] Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands

[4] Department of Psychology, Utrecht University, Utrecht, The Netherlands

[5] Cito Institute for Educational Testing, Arnhem, The Netherlands

replications with a dichotomous outcome (0 = failure, 1 = success) in 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 $\boldsymbol{\pi}^i = [\pi_1^i, \ldots, \pi_J^i]$. 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 0.5, however this does not imply that each individual has a penalty scoring probability smaller than 0.5. 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.

P-population and WP-population

An example of a within-subject experiment is Zedelius, Veling, and 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 (17 ms) or *supraliminally* (sup), that is for a longer duration of 300 ms. 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

$$\begin{aligned} H_1 : \textit{hr-sup-li} > \textit{hr-sup-hi} > \textit{hr-sub-li} > \textit{hr-sub-hi} \\ > \textit{lr-sup-li} > \textit{lr-sup-hi} > \textit{lr-sub-li} > \textit{lr-sub-hi}, \end{aligned} \tag{1}$$

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:

$$\begin{aligned} H_1^i : \textit{hr-sup-li}^i > \textit{hr-sup-hi}^i > \textit{hr-sub-li}^i > \textit{hr-sub-hi}^i \\ > \textit{lr-sup-li}^i > \textit{lr-sup-hi}^i > \textit{lr-sub-li}^i > \textit{lr-sub-hi}^i, \end{aligned} \tag{2}$$

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

To illustrate the difference between Eqs. 1 and 2 let us consider a *population of persons* (P-population from here on) 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 1 is a hypothesis regarding the ordering of these P-population probabilities. Equation 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.

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.

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

$$H_m^i : R_m\boldsymbol{\pi}^i > 0, \tag{3}$$

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:

$$R_1 = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \tag{4}$$

which renders

$$H_1^i : \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i, \tag{5}$$

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

$$R_2 = \left[\ .5\ .5\ -.5\ -.5\ \right], \tag{6}$$

renders the informative hypothesis

$$H_2^i : \frac{\pi_1^i + \pi_2^i}{2} > \frac{\pi_3^i + \pi_4^i}{2}, \tag{7}$$

which states that the average of the first two success probabilities is larger than the average of the last two. Hypotheses constructed using Eq. 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:

$$H_{\cancel{m}}^i : \text{not } H_m^i. \tag{8}$$

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:

$$H_u^i : \pi_1^i, \pi_2^i, \ldots, \pi_{J-1}^i, \pi_J^i, \tag{9}$$

where each parameter is 'free'. An informative hypothesis H_m^i constrains the parameter space such that only particular combinations of parameters are allowed, $H_{\cancel{m}}^i$ comprises that part of the parameter space that is not included in H_m^i and the conjunction of H_m^i and $H_{\cancel{m}}^i$ is H_u^i. The difference in use of H_u^i and $H_{\cancel{m}}^i$ 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., 2014). A first translation of the expectations is

$$\begin{aligned} H_1^i : \textit{hr-sup-li}^i &> \textit{hr-sup-hi}^i > \textit{hr-sub-li}^i > \textit{hr-sub-hi}^i \\ &> \textit{lr-sup-li}^i > \textit{lr-sup-hi}^i > \textit{lr-sub-li}^i > \textit{lr-sub-hi}^i, \end{aligned} \tag{10}$$

which states that for any person i the success probabilities are ordered from high to low. To give some intuition for this hypothesis, Fig. 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 Fig. 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:

$$H_2^i : hr\text{-}li^i > hr\text{-}hi^i > lr\text{-}li^i > lr\text{-}hi^i\,, \tag{11}$$

and

$$H_3^i : hr\text{-}sup^i > hr\text{-}sub^i > lr\text{-}sup^i > lr\text{-}sub^i\,, \tag{12}$$

where *hr-li*^{i} indicates the average success probability of the *hr-sup-li*^{i} and *hr-sub-li*^{i} conditions. In Fig. 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 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:

$$H_4^i : hr\text{-}sup^i - lr\text{-}sup^i > hr\text{-}sub^i - lr\text{-}sub^i\,, \tag{13}$$

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 Fig. 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.

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 $\mathbf{x}^i = [x_1^i, \ldots, x_J^i]$ observed

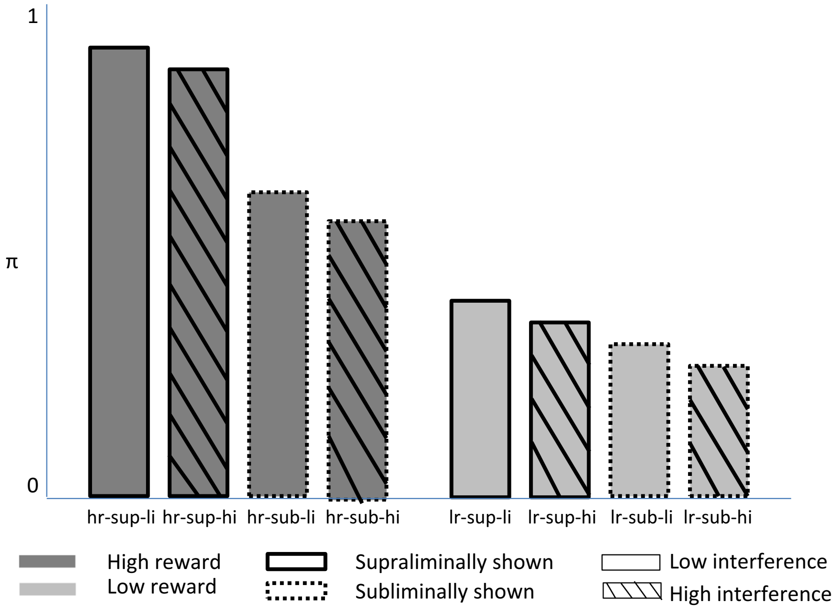

Fig. 1 Graphical representation of all hypotheses by Zedelius et al. (2011)

for person i in R replications in each condition j the density of the data is

$$f(\mathbf{x}^i \mid \boldsymbol{\pi}^i) = \prod_{j=1}^{J} \binom{R}{x_j^i} (\pi_j^i)^{x_j^i} (1 - \pi_j^i)^{R - x_j^i}, \tag{14}$$

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

$$h(\boldsymbol{\pi}^i \mid H_u^i) = \prod_{j=1}^{J} \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} (\pi_j^i)^{\alpha_0 - 1} (1 - \pi_j^i)^{\beta_0 - 1}, \tag{15}$$

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, Laudy, & Hoijtink, 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:

$$\begin{aligned} g(\boldsymbol{\pi}^i \mid \mathbf{x}^i, H_u^i) &\propto f(\mathbf{x}^i \mid \boldsymbol{\pi}^i) \cdot h(\boldsymbol{\pi}^i \mid H_u^i) \\ &\propto \prod_{j=1}^{J} \frac{\Gamma(\alpha_1 + \beta_1)}{\Gamma(\alpha_1)\Gamma(\beta_1)} (\pi_j^i)^{\alpha_1 - 1} (1 - \pi_j^i)^{\beta_1 - 1}, \end{aligned} \tag{16}$$

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 Eq. 16, the posterior distribution is indeed only dependent on the data.

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, p. 51–52, 57–59) show that for inequality constrained hypotheses of the form presented in Eq. 3 the ratio of marginal likelihoods expressing support for H_m^i relative to H_u^i can be rewritten as

$$BF_{mu}^i = \frac{f_m^i}{c_m^i}. \tag{17}$$

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

$$c_m^i = \int_{\boldsymbol{\pi}^i \in H_m^i} h(\boldsymbol{\pi}^i \mid H_u^i) \delta \boldsymbol{\pi}^i. \tag{18}$$

Using Eq. 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 Eq. 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:

$$f_m^i = \int_{\boldsymbol{\pi}^i \in H_m^i} g(\boldsymbol{\pi}^i \mid \mathbf{x}^i, H_u^i) \delta \boldsymbol{\pi}^i. \tag{19}$$

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 Eq. 17 is a ratio of two marginal likelihoods (one for H_m^i and one for H_u^i) it follows that

$$BF_{mm'}^i = \frac{BF_{mu}^i}{BF_{m'u}^i} = \frac{f_m^i / c_m^i}{f_{m'}^i / c_{m'}^i}, \tag{20}$$

and that

$$BF_{m\not{m}}^i = \frac{f_m^i / c_m^i}{f_{\not{m}}^i / c_{\not{m}}^i} = \frac{f_m^i / c_m^i}{1 - f_m^i / 1 - c_m^i}. \tag{21}$$

Three hypothetical $N = 1$ datasets with $J = 4$ and $R = 7$ are presented in Table 1. Three possible informative hypotheses regarding these data are H_1^i from Eq. 5, $H_{\not{1}}^i$ and H_2^i from Eq. 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 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

Table 1 Complexity, fit, and Bayes factors for three hypothetical $N = 1$ studies with $H_1^i = \pi_1^i > \pi_2^i > \pi_3^i > \pi_4^i$ and $H_2^i = \frac{\pi_1^i+\pi_2^i}{2} > \frac{\pi_3^i+\pi_4^i}{2}$

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	$BF_{1\not{1}}^i$	BF_{12}^i	$BF_{2\not{2}}^i$
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

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 for H_m^i and H_u^i caused by the data. For example, in Table 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}^i 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}^i and $BF_{m\not{m}}^i$ are always both larger or smaller than 1. However, by definition BF_{mu}^i ranges from 0 to $\frac{1}{c_m^i}$ and $BF_{m\not{m}}^i$ 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 $BF_{m\not{m}}^i$, 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 1, $BF_{1\not{1}}^1 = 28.4$ indicates that the data caused an increase in believe for H_m^i compared to $H_{\not{m}}^i$, 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 (e.g. Kass & Raftery, 1995, interpret 3 as the demarcation for the size of BF_{ab}, providing marginal and positive evidence in favor of H_a), we choose not to follow them. In the spirit of a famous quote from Rosnow and 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 two 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 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 $BF_{1\not{1}}^2 = 1.43$ and $BF_{2\not{2}}^2 = 8.09$, 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 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, $BF_{2\not{2}}^3 = 1.04$, indicating that although H_2^i is better than H_1^i, it is not a very good hypothesis. The examples in Table 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.

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 Eqs. 10, 11, 12 and 13. The Bayes factors presented in the first four columns of Table 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 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 $BF^8_{2\not{2}} = .15$ and $BF^8_{4\not{4}} = .10$ we

Table 2 Individual Bayes factors for the Zedelius data where H_1^i, H_2^i, H_3^i and H_4^i (Eqs. 10–13) are evaluated against H_u^i and their complement

i	BF^i_{1u}	BF^i_{2u}	BF^i_{3u}	BF^i_{4u}	$BF^i_{1\not{1}}$	$BF^i_{2\not{2}}$	$BF^i_{3\not{3}}$	$BF^i_{4\not{4}}$
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

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 $BF_{2\not{2}}^{14} = 8.61$, 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 $BF_{4\not{4}}^{20} = .65$ 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 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 2 is that this is not the case. However, this topic will be pursued in depth in the next section.

A P-population of WP-populations

Looking at the Bayes factors in Table 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 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):

$$\text{P-BF}_{mu} = \prod_{i=1}^{P} \text{BF}_{mu}^{i}, \tag{22}$$

which expresses the support for H_m relative to H_u, where

$$H_m = H_m^1 \cup \ldots \cup H_m^P, \tag{23}$$

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

$$H_u = H_u^1 \cup \ldots \cup H_u^P, \tag{24}$$

which is the union of H_u^i for $i = 1, \ldots, P$. 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 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_m^i. Thus, it is a measure of the extent to which a hypothesis holds for every individual, rather than on average.

Table 3 shows seven hypothetical sets of six individual Bayes factors comparing H_m^i to H_u^i. 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_m^i 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 and Penny (2007) have suggested using the geometric mean of the product of individual Bayes factors to render a summary that is independent of P:

$$\text{gP-BF}_{mu} = \sqrt[P]{\text{P-BF}_{mu}}, \tag{25}$$

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 $P + 1^{\text{st}}$ individual sampled from the P-population.

As can be seen in Table 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 $BF_{m\not{m}}$ can be derived similar as $BF^{i}_{mm'}$ and $BF^{i}_{m\not{m}}$. 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 $BF^{i}_{m\not{m}}$ are always both smaller or larger than 1. For example, if $BF^{1}_{1u} = 0.2$, then $BF^{1}_{1\not{1}} = 0.4$, and if $BF^{2}_{1u} = 1.8$ then $BF^{2}_{1\not{1}} = 9$. This is not true for gP-BF_{mu} and gP-$BF_{m\not{m}}$. To continue the example based on the Bayes factors for persons 1 and 2, gP-$BF_{1u} = 0.6$ and gP-$BF_{1\not{1}} = 2$. 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-$BF_{m\not{m}}$, which is a summary of all $BF^{i}_{m\not{m}}$. 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 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:

$$ER_{mu} = \begin{array}{ll} \frac{1}{P}\sum_{i=1}^{P} I_{BF^{i}_{mu}<1} & \text{if gP-BF}_{mu} < 1 \\ \frac{1}{P}\sum_{i=1}^{P} I_{BF^{i}_{mu}>1} & \text{if gP-BF}_{mu} > 1 \end{array}, \qquad (26)$$

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, 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:

$$SR_{mu} = \begin{array}{ll} \frac{1}{P}\sum_{i=1}^{P} I_{BF^{i}_{mu}<\text{gP-BF}_{mu}} & \text{if gP-BF}_{mu} < 1 \\ \frac{1}{P}\sum_{i=1}^{P} I_{BF^{i}_{mu}>\text{gP-BF}_{mu}} & \text{if gP-BF}_{mu} > 1 \end{array}, \qquad (27)$$

where $I_{BF^{i}_{mu}<\text{gP-BF}_{mu}} = 1$ if $BF^{i}_{mu} <$ 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 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 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

Table 3 Hypothetical individual Bayes factors ($P = 6$), gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ underlined entries indicate individual Bayes factors smaller than the gP-BF and bold entries indicate entries larger than the corresponding gP-BF

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
BF^1_{mu}	**9.00**	**3.20**	1.40	$\underline{0.80}$	$\underline{0.90}$	**6.40**	1.01
BF^2_{mu}	7.11	2.70	**2.70**	1.50	$\underline{0.93}$	1.40	1.30
BF^3_{mu}	–	2.30	1.80	**2.50**	$\underline{0.85}$	1.80	**2.50**
BF^4_{mu}	–	**3.22**	**2.10**	**4.33**	$\underline{0.88}$	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

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 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-BF$_{mm'}$ or gP-BF$_{m\not{m}}$, the ER and SR should be computed using the individual $BF^i_{mm'}$s and $BF^i_{m\not{m}}$s. The individual Bayes factors are the relevant quantities in the ER and SR, and therefore these should be used.

Illustration

Using the individual Bayes factors presented in Table 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 4 gives the gP-BF$_{mu}$ based on the individual Bayes factors from Table 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$_{2\not{2}} = 1.014$, gP-BF$_{3\not{3}} = 1.235$ and gP-BF$_{4\not{4}} = 1.412$, 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 neuropsychological research. The theory and hypotheses for these

Table 4 The gP-BF, ER and SR for the data of Zedelius et al. (2011) for the evaluation of H_1^i, H_2^i, H_3^i and H_4^i (Eqs. 10–13)

	BF_{1u}	BF_{2u}	BF_{3u}	BF_{4u}	$BF_{1\not{1}}$	$BF_{2\not{2}}$	$BF_{3\not{3}}$	$BF_{4\not{4}}$
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

experiments are often at the WP-population level. Examples are Moreland and 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.

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 14, 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 version 3.3.1 (R Core Team, 2013), software has been developed with which such a study design analysis can be executed.[1] 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_m^i 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

[1] 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 assistance with or questions about the software, please also contact the first author.

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:

- 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.

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, Eqs. 10–12, about the ordering of success probabilities in the $J = 8$ experimental conditions. For this illustration, only H^i_1 as in Eq. 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^i_1. From this hypothesis, three relevant P-populations are derived.
  - *P-population 1.* In this P-population all individuals adhere to H^i_1. Using this population the median values of the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ can be determined if H^i_m holds for everyone. To compute these median values, the individual parameters $\boldsymbol{\pi}^i$ are repeatedly sampled from the prior distribution under H^i_1:

$$h(\boldsymbol{\pi}^i|H^i_1) \propto h(\boldsymbol{\pi}^i|H^i_u) I_{\boldsymbol{\pi}^i\in H^i_1}, \tag{28}$$

    where $I_{\boldsymbol{\pi}^i\in H^i_1} = 1$ if $\boldsymbol{\pi}^i$ is in agreement with H^i_1 and 0 otherwise.
  - *P-population 2.* In this P-population all individuals adhere to $H_{\not 1}$. Using this population, the expected values of the gP-BF$_{mu}$, ER$_{mu}$ and SR$_{mu}$ can be determined if $H^i_{\not m}$ are sampled from the prior distribution under $H^i_{\not 1}$, that is:

$$h(\boldsymbol{\pi}^i|H^i_{\not 1}) \propto h(\boldsymbol{\pi}^i|H^i_u) I_{\boldsymbol{\pi}^i\in H^i_{\not 1}}, \tag{29}$$

    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 Eq. 28 if u^i, sampled from

$U(0, 1)$ is smaller than or equal to the specified proportion θ, and sampled from Eq. 29 if u^i is larger than θ:

$$\pi^i \sim \begin{array}{l} h(\pi^i|H_1^i) \text{ if } u^i \leq \theta \\ h(\pi^i|H_{\not 1}^i) \text{ if } u^i > \theta \end{array}. \qquad (30)$$

The proportion θ is set to .5, thus half of all individuals adheres to H_1^i and the other half adheres to $H_{\not 1}^i$.

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 seven 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.

Results

Figure 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_1^i 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_1^i holds for all

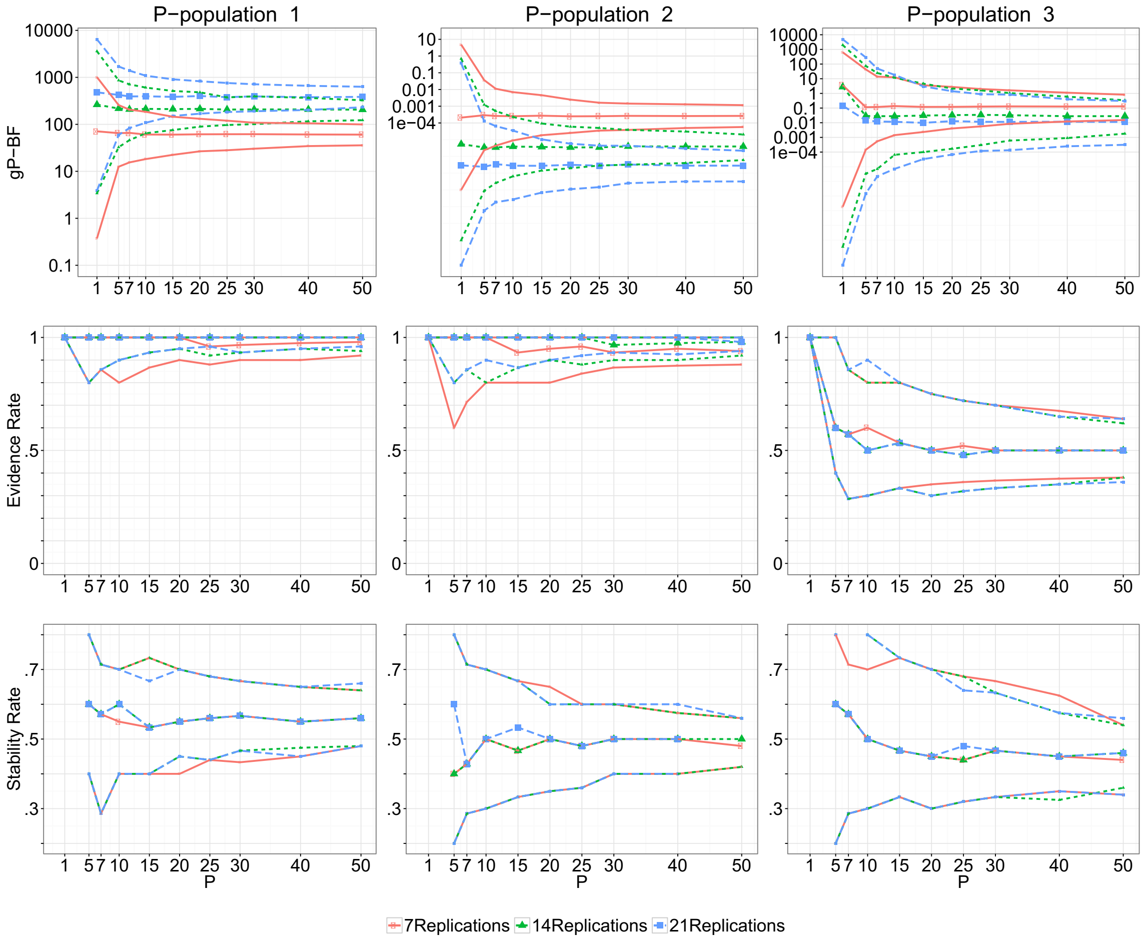

Fig. 2 gP-BF$^i_{1\not1}$, ER$^i_{1\not1}$ and SR$^i_{1\not1}$ for the three generated true P-populations for $J = 8$. P-population 1 is described in Eq. 28, P-population 2 in Eq. 29, and P-population 3 in Eq. 30. Both the median and 95% interval are shown in the figures

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^i_1 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 Fig. 2 $H^i_{\not 1}$ is true for all individuals. For $P > 10$ and $R > 7$, the gP-BF is smaller than .01, indicating at least ten times more support for $H^i_{\not 1}$ than for H^i_1. 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 Fig. 2 was chosen to be a mixture of the first two populations. Here it can be seen that if H^i_1 holds for 50% of the individuals in the population, generally, $H^i_{\not 1}$ is preferred over H^i_1, 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 $h(\boldsymbol{\pi}^i|H^i_1)$ provides evidence in agreement with $H^i_{\not 1}$ 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 $H_{\tilde{A}\breve{g}ancel1}$ is $1 - 2.48 \times 10^{-5} \approx 1$. Thus, even though $\theta = .5$, $H^i_{\not 1}$ is preferred because it has a higher complexity. The $\text{ER}_{m\not m}$ 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-$\text{BF}_{m\not m}$ might indicate a preference for $H^i_{\not 1}$ over H^i_1, the $\text{ER}_{m\not m}$ indicates inconsistency among individual Bayes factors. Finally, the median $\text{SR}_{m\not m}$ 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 $\text{ER}_{m\not m}$ for every P and R that seemed reasonable if H^i_1 or $H^i_{\not 1}$ were true for everyone.

Zedelius et al. (2011) eventually used 26 participants in their study and seven replications. This is slightly lower than the suggested 30 based on the sensitivity analysis. Consulting the figures, it seems that, if H^i_1 is true and $P = 26$ and $R = 7$, gP-$\text{BF}_{1\not 1}$ is expected to be between 30 and 100, the $\text{ER}_{1\not 1}$ is expected to be above .9 and the $\text{SR}_{1\not 1}$ between .43 and .67. On the other hand, if $H^i_{\not 1}$ is true for all individuals, the gP-$\text{BF}_{1\not 1}$ can be expected between 1000 and 10,000 in support of $H^i_{\not 1}$, with the $\text{ER}_{1\not 1}$ similarly above .9 and the $\text{SR}_{1\not 1}$ between .35 and .6. Consulting the results in Table 4, we find that gP-$\text{BF}_{1\not 1} = .511$, $\text{ER}_{mu} = .500$ and $\text{SR}_{mu} = .436$. These results do not seem in line with either Population 1 or 2, but consulting the right column figures in Fig. 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-$\text{BF}_{1\not 1}$ shows some support for $H^i_{\not 1}$ relative to H^i_1, it is not likely that $H^i_{\not 1}$ holds for everyone in the P-population.

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.

Author Note This research was supported by a grant from The Netherlands Organisation for Scientific Research (NWO): NWO 406-12-001 (FK and HH). The last author is supported by the Consortium on Individual Development (CID) which is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003). The starting point for this paper was the Master thesis by Kluytmans et al. (2014).

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.

Appendix: Computation of fit and complexity through decomposition

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

$$H_m : R_m\boldsymbol{\pi} > 0 \tag{31}$$

against the unconstrained hypothesis H_u, the complexity and fit of H_m should be computed.[2]

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

$$\begin{aligned} f_m &= \int_{\boldsymbol{\pi} \in H_m} g(\boldsymbol{\pi}|\mathbf{x}, H_u)\delta\boldsymbol{\pi} \\ &\approx \frac{1}{Q}\sum_{q=1}^{Q} I_{\boldsymbol{\pi}^q \in H_m} \end{aligned} \tag{32}$$

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 eight 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, Hoijtink, and de Leeuw (2012) present another solution that makes use of a decomposition of the complexity and fit. This procedure determines decomposed fit and complexity for each constraint in a hypothesis. Equation 33 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:

$$\begin{aligned} P(\mathbf{R}_m\boldsymbol{\pi} > 0|H_u, \mathbf{x}) &= \prod_{k=1}^{K} P(\mathbf{R}_m^{(k)}\boldsymbol{\pi} > 0|H_u, \mathbf{x}, \mathbf{R}_m^{(1:k-1)}) \\ &= \prod_{k=1}^{K} f_m^{(k)} \\ &\approx \prod_{k=1}^{K} \frac{1}{Q}\sum_{q=1}^{Q} I_{\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q > 0}, \end{aligned} \tag{33}$$

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 Eq. 33 the total fit can be obtained accurately.

The complexity can be derived analogously:

$$\begin{aligned} P(\mathbf{R}_m\boldsymbol{\pi} > 0|H_u) &= \prod_{k=1}^{K} P(\mathbf{R}_m^{(k)}\boldsymbol{\pi} > 0|H_u, \mathbf{R}_m^{(1:k-1)}) \\ &= \prod_{k=1}^{K} c_m^{(k)} \\ &\approx \prod_{k=1}^{K} \frac{1}{Q}\sum_{q=1}^{Q} I_{\mathbf{R}_m^{(k)}\boldsymbol{\pi}^q > 0}, \end{aligned} \tag{34}$$

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

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

References

Hamaker, L. E. (2012). Mehl, M. R., & Conner, S. (Eds.) *Handbook of research methods for studying daily life*, (pp. 43–61). New York: Guilford.

Hoijtink, H. (2012). *Informative hypotheses. Theory and practice for behavioral and social scientists*. Boca Raton: Chapman & Hall/CRC.

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

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. 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. https://doi.org/10.1037/1082-989X.10.4.477

Kluytmans, A., Van de Schoot, R., Zedelius, C., Veling, H., Aarts, H., & Hoijtink, H. (2014). *Bayesian sequential evaluation of simple order constraints using dichotomous within-subject data*. Unpublished manuscript.

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. https://doi.org/10.1080/19312458.2015.1096332.

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. https://doi.org/10.1207/s15366359mea0204_1

Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal of Experimental Social Psychology*, *18*, 395–415. https://doi.org/10.1016/0022-1031(82)90062-2.

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

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria. http://www.R-project.org/. Accessed date: 24 January 2017.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, *44*(10), 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Stephan, K. E., & Penny, W. D. (2007). Dynamic causal models and Bayesian selection. In Friston, K., Ashbumer, J., Kievel, S., Nichols, T., & Penny, W. (Eds.) *Statistical parametric mapping: The analysis of functional brain images* (pp. 577–585): Academic Press.

Woodcock, J. (2007). The prospects for “personalized medicine” in drug development and drug therapy. *Clinical Pharmacology and Therapeutics*, *81*(2), 164–169. https://doi.org/10.1038/sj.clpt.6100063

Zedelius, C. M., Veling, H., & Aarts, H. (2011). Boosting or choking – how conscious and unconscious reward processing modulate the active maintenance of goal-relevant information. *Consciousness and Cognition*, *20*, 355–362. https://doi.org/10.1016/j.concog.2010.05.001