Embracing uncertainty in multi-step inference

van den Bergh, D.

Publication date
2024

Document Version
Final published version

Link to publication

Citation for published version (APA):
van den Bergh, D. (2024). *Embracing uncertainty in multi-step inference*. [Thesis, fully internal, Universiteit van Amsterdam].

In textbooks, we often come across studies where a perfect data set is analyzed with a statistical test that tells us everything we want to know. However, a typical day in the life of a scientist could not be any more different. The data are messy, the research questions are not easily translated into statistical tests and there is no single best statistical test. In practice, the next best thing is to break the statistical analysis up into multiple steps. By conducting multiple statistical tests, each informed by and building on the previous one, we hope to meaningfully answer a complex and otherwise unanswerable research question. This multi-step procedure, however, has a fatal flaw. The result of any statistical test is subject to some uncertainty. Subsequent analyses should account for this uncertainty, otherwise the results are prone to overconfidence.

This dissertation focuses on the uncertainty that arises when conducting multi-step inference and consists of three parts. The first part discusses uncertainty within statistical models and shows that decisions that precede an analysis may already conceal uncertainty. The second part discusses uncertainty between statistical models and how this can be handled through Bayesian model averaging. The final part focuses on making the methodology in the previous parts easily and freely accessible by implementing it in the free open-source statistical software program JASP.

Embracing Uncertainty in Multi-Step Inference

Don van den Bergh

Embracing Uncertainty in Multi-Step Inference

Don van den Bergh

Embracing Uncertainty in Multi-Step Inference

Don van den Bergh

ISBN 978-94-6469-748-3
This thesis was typeset with LaTeX using the Dissertate class developed by Jordan Suchow and adjusted by Don van den Bergh.
Printed by ProefschriftMaken || www.proefschriftmaken.nl
Cover design by Ron Zijlmans || www.ron.nu

The research of this doctoral thesis received financial assistance from a Research Talent grant from the Netherlands Organisation of Scientific Research (NWO).

Embracing Uncertainty in Multi-Step Inference

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.-P. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdagdag 13 maart 2024, te 16:00 uur

door

Don van den Bergh

geboren te Amsterdam

Promotiecommissie

Promotor:	Prof. dr. E. M. Wagenmakers	Universiteit van Amsterdam
Copromotores:	dr. M. Marsman	Universiteit van Amsterdam
	dr. A. Ly	Universiteit van Amsterdam
Overige leden:	prof. dr. M. A. Clyde	Duke University
	prof. dr. H. L. J. van der Maas	Universiteit van Amsterdam
	prof. dr. D. Borsboom	Universiteit van Amsterdam
	dr. Z. Oravecz	Pennsylvania State University
	dr. ir. J. Mulder	Universiteit Tilburg
	dr. D. Matzke	Universiteit van Amsterdam
	dr. J. Haaf	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragswetenschappen

University of Amsterdam

Contents

Absolute certainty is a privilege of uneducated minds and fanatics.
It is, for scientific folk, an unattainable ideal.

Cassius J. Keyser

1
Embracing Uncertainty

A characteristic feature of empirical science is dealing with uncertainty. Uncertainty usually stems from many sources and finds its way into empirical research in both expected and unexpected ways. For example, suppose a researcher intends to measure a participant's mental well-being with a questionnaire. They ask the participant a series of questions and the participant responds honestly. In this situation, the participant's responses, that is, the data, are noisy measurements of what the researcher is interested in, that is, the participant's well-being. To account for this uncertainty, researchers typically ask multiple questions and assume that the noise in answers to individual questions averages out. This is a well-known source of uncertainty called measurement error. A lesser-known source of uncertainty originates from statistical models and the decisions made in the interim when conducting statistical analyses. For example, statistical models, and the corresponding analyses, often make assumptions. These assumptions, however, may be violated and therefore the conclusions of those analyses hinge on the (un)certainty with which their assumptions are met. In empirical practice, we often see that this results in a two-step procedure. First, a test is conducted to examine whether the assumptions made by a subsequent test are violated. If these assumptions are not violated, then the test of interest is executed. However, this two-step procedure has a subtle flaw. The uncertainty reported by the second test does not account for the (un)certainty with which the first test determined if the assumptions were violated. In fact, the uncertainty associated with the first test is completely ignored, causing the second test to produce overconfident results. More generally, the data analysis in empirical studies often consists of multiple intermediate steps and interim decisions. Each of these steps typically involves uncertainty and virtually no intermediate decision is made with absolute certainty. In many studies, however, the uncertainty of the various intermediate steps is ignored, creating a false level of confidence that can affect the conclusions. The central theme of this dissertation is the uncertainty that arises when conducting multi-step inference and how to incorporate all uncertainty into the final results.

1.1 Uncertainty Within a Single Model

The first part of this dissertation concerns itself with uncertainty *within* a statistical model. This part illustrates that decisions that precede the application of a statistical model or analysis may conceal uncertainty. Suppose a group of patients with a mental disorder is scored by five psychiatrists on a variety of items, such as psychotic behavior, impulsive behavior, and problem insight. A common first step in analyzing such data is to take the sample mean of all the five psychiatrists to obtain one score for each patient on each item. In a second step, these averages are admitted to some statistical test. This two-step procedure ignores an important source of uncertainty, namely that the different psychiatrists did not all give the same score. More specifically, naively averaging makes an implicit assumption that the five psychiatrists are exchangeable and that their individual differences are irrelevant. Typically, this assumption is unwarranted because different psychiatrists have different backgrounds and may have meaningful individual differences that result in heterogeneous scoring behavior. Nevertheless, by sweeping this source of uncertainty under the rug, the variability in the responses is neglected. This leads to overconfidence in what is now the "observed" data (i.e., the average scores) that are fed into subsequent analyses. To see why this leads to overconfidence, note that, at a minimum, this procedure ignores the standard errors associated with the sample means. Therefore, the observations used by subsequent analyses are in fact more variable than the analyses are aware of. However, because all subsequent analyses are blind to this variability, the uncertainty intervals are narrower than they should be, leading to overconfident conclusions.

The aim of the first part of this dissertation was to develop models that explicitly account for the fact that different individuals gave the scores and to properly quantify the uncertainty. While the data analyzed in this part focuses on patients in forensic psychiatric hospitals and is therefore quite specialized, the structure of the data analyzed is rather common in empirical science. For example, data obtained through questionnaires that are repeatedly filled out by participants or data in educational psychology where essays or other products are scored by multiple raters tend to have a similar structure. As such, the methods developed in this part can be generalized to other applications outside of a forensic setting.

1.2 Uncertainty Between Multiple Models

The second part of this dissertation concerns itself with uncertainty when multiple models are in play. Suppose we want to predict the risk of a violent outburst in the cohort of our mentally ill patients. Using all available data, we want to construct a model that accurately predicts the risk of a violent outburst in the future. However, if we naively include all items scored by psychiatrists and other background variables, we run the risk of overfitting and our model may poorly predict future violent outbursts. A common approach to combat overfitting is to use a two-step procedure where in the first step a single model is selected and in the second step that model is interpreted and used for predictions. However, this two-step procedure again systematically ignores uncertainty uncertainty and can lead to overconfident conclusions. There is considerable uncertainty about which model is the "best" model that should be used in the second step. Sometimes, there is no single best model superior to all other candidate models. Instead, there often are a plethora of models that make adequate predictions and offer reasonable explanations for the data. Nevertheless, by selecting a single model, we ignore this model uncertainty and proceed as if we have discovered the one true model to base our inferences on. As a result, we become overconfident and overestimate the size of the established effects (e.g., Hoeting et al., 1999; Porwal and Raftery, 2022, Chapter 5).

The aim of the second part of this dissertation was to develop new statistical methods to quantify uncertainty across different models. The key approach that is central to this part is *model averaging*. Rather than selecting a single model for prediction, we make predictions using all models considered and weigh the predictions of each model by its relative plausibility in light of the data.

1.3 Embracing Uncertainty for Everyone

The third part of this dissertation is about making model averaging accessible to practitioners without a mathematical background and programming knowledge. A large part of the statistical literature is concerned with the development of novel and important methods. However, the road for those who want to put these methods into practice is usually full of obstacles, if a clear path can be seen at all. For example, practitioners may lack the mathematical background to understand the derivations or the programming expertise to implement a new technique. These barriers make it difficult to put new developments into practice. The third part reflects on the literature on Bayesian model averaging and focused on adapting, fine-tuning, and developing model averaging techniques to statistical paradigms relevant for psychological practice. This was done by implementing the techniques in the free and open-source statistical software program JASP (JASP Team, 2022). As a result practitioners can bring the ideas and techniques developed in the first two parts of the dissertation into practice without the need for an in-depth mathematical background or advanced programming knowledge.

1.4 Chapter outline

1.4.1 Cultural Consensus Theory

The first part of this dissertation concerns itself with uncertainty within a model. Specifically, it provides an alternative to the common practice of averaging the sample scores of different raters into a single value by using approaches based on cultural consensus theory and signal detection theory.

In Chapter 2, we discuss a parsimonious approach to estimating latent thresholds and apply this to a signal detection theory model. This chapter exemplifies a trade-off in uncertainty in signal detection theory models for ordinal data. The usual approach requires large amounts of data to estimate the threshold parameters and may result in overly complex models. In contrast, our proposed approach makes stronger assumptions but as a consequence the model parameters are more readily estimable. In Chapter 3, we develop a model based on cultural consensus theory to analyze data from patients in forensic psychiatric hospitals. We reuse the parsimonious approach for estimating thresholds from chapter one. Chapter 4 applies the model developed in Chapter 2 to data from patients in the Dutch maximum-security Forensic Psychiatric Center Dr. S. van Mesdag. We use the cultural consensus theory model developed in Chapter 3 to augment a logistic regression model with the aim of predicting violent outbursts in patients.

1.4.2 Bayesian Model Averaging

The second part of this dissertation discusses uncertainty when multiple models are considered. It is very common for researchers to first use some procedure to decide on a single model, and then in a second step draw conclusions conditional on the selected model. However, when drawing conclusions, all uncertainty about the models is ignored (Hinne et al., 2020). This results in overconfident conclusions.

Chapter 5 illustrates this overconfidence in the simple case of estimating an effect size for a t-test. We illustrate that the common practice of first conducting a significance test to reject the null hypothesis and afterward using the alternative hypothesis as the ground truth yields overestimated effect sizes. In Chapter 6, we develop a confirmatory default Bayesian hypothesis test for comparing the variances of multiple variables with mixed equality and inequality constraints. The proposed approach is confirmatory in the sense that the models to be compared must be selected beforehand. As such our method may lead scientists to neglect model uncertainty, if they use our confirmatory test while they in reality are unsure about which models to compare at all. To address this shortcoming, Chapter 7 extends Chapter 6 and introduces an exploratory method that considers all possible equality constraints among the variances of multiple variables. Furthermore, Chapter 7 explores two families of prior distributions that can be used to penalize models with different equality constraints in different ways. In addition, we introduce a general method that goes beyond comparing variances and that can be used when testing for equality constraints among any parameter vector. For example, we provide data examples where we test equality constraints among means and proportions.

1.4.3 JASP

The third part of this dissertation is of a more pragmatic nature and strives to improve the accessibility of advanced statistical methods. Even if advanced techniques are developed to embrace uncertainty and avoid overconfidence, they are of little use if they are not easily accessible for practitioners. Therefore, the chapters in this part provide tutorials on Bayesian model averaging and implementations thereof in the free open-source statistical software program JASP.

Chapter 8 provides a tutorial on Bayesian multi-model linear regression. We explain the theory behind linear regression, Bayesian inference, and Bayesian multi-model inference. Afterward we illustrate Bayesian multi-model linear regression on a data example about happiness scores of different countries. In a similar fashion, Chapter 9 demonstrates how to conduct a Bayesian Analysis of Variance (ANOVA) while accounting for model uncertainty. We use two data examples; one to show how to interpret the general results, and another to demonstrate how to conduct post-hoc tests. Chapter 10 builds on Chapter 9 and provides a more thorough foundation for the choice of model space, that is, the space of all candidate models under consideration. We argue that the most commonly used model space approach up to this point is suboptimal for repeated-measures ANOVA because it implies that there are no individual differences in the effects. Instead, we propose an alternative model space that always models individual differences and motivate the decision to make this the default model space in JASP.

Part I

Cultural Consensus Theory

1

2
Parsimonious Estimation of Signal Detection Models from Confidence Ratings

Signal Detection Theory (SDT) is used to quantify people's ability and bias in discriminating stimuli. The ability to detect a stimulus is often measured through confidence ratings. In SDT models, the use of confidence ratings necessitates the estimation of confidence category thresholds, a requirement that can easily result in models that are overly complex. As a parsimonious alternative, we propose a threshold SDT model that estimates these category thresholds using only two parameters. We fit the model to data from Pratte et al. (2010) and illustrate its benefits over previous threshold SDT models.

This chapter is published as: Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*, 1953–1967.

Our ability to recognize stimuli allows us to interact smoothly with the world. We know that if we want to drink water it is a good idea to poor it into a cup instead of onto a piece of paper. We also know that if we want to write something down it is a good idea to use a pen instead of a yoga mat. Although recognizing stimuli is sometimes straightforward, often it is not. Most of the times, our ability to recognize a stimulus is accompanied by a certain amount of noise. When picking mushrooms it can be hard to distinguish between the mushrooms you can use to top your beautiful saffron risotto, and the mushrooms that will turn your dinner party into the next Jonestown. Not only do eatable and poisonous mushrooms differ in perceptual similarity—it is easy to classify a mushroom with a red cap and white spots as poisonous, but difficult to do so for a poisonous mushroom that looks similar to a common white button mushroom—but the amount of risk involved in making the wrong decision can also differ between situations: when you are starving you might decide to eat a suspicious looking mushroom sooner than when you just had a full course meal. Signal Detection Theory (SDT; Green & Swets, 1966; Tanner Jr. & Swets, 1954) disentangles these aspects of recognition by providing different parameters: (1) the amount of information that is available in the stimulus, and (2) the threshold you set for making one or the other decision.

In order to separately estimate these two aspects of recognition, an SDT model needs two pieces of information: (1) the proportion of correctly identified signal stimuli (hit rate, HR; the proportion of poisonous mushrooms that were correctly identified as poisonous), and (2) the proportion of incorrectly identified noise stimuli (false alarm rate, FAR; the proportion of non-poisonous mushrooms that were incorrectly identified as poisonous). Table 2.1 depicts the four possible outcomes when discriminating two types of stimuli; Equation 2.1 and 2.2 show how these outcomes can be converted to hit rate and false alarm rate:

Table 2.1: Possible outcomes when trying to discriminate signal from noise stimuli. The rows represent the estimates and the columns represent the truth.

		Truth	
		Signal	Noise
Response	Signal	Hit	False Alarm
	Noise	Miss	Correct Rejection

$$\text{Hit Rate} = \frac{\text{Hits}}{\text{Hits} + \text{Misses}}, \tag{2.1}$$

$$\text{False Alarm Rate} = \frac{\text{False Alarms}}{\text{False Alarms} + \text{Correct Rejections}}. \tag{2.2}$$

SDT is a popular model for the analysis of experiments in recognition memory. The most common experiment in this field first requires that participants study a list of words (i.e., the study list). Following a retention interval, participants are presented with another (i.e., the test list, containing words from the study list and new words). For each word on the test list, participants are asked to decide whether the word was from the study list (i.e., 'old'), or not (i.e., 'new'). Figure 2.1 illustrates how the SDT model uses hit and false alarm rates to identify the strength of the signal, d', in this task and the threshold, λ, that is set to make one or the other decision. To estimate these two parameters, the model assumes that both the signal (i.e., 'old' words) and the noise (i.e., 'new' words) stimuli can be placed on a latent continuous scale of familiarity. The latent scores are drawn from a signal normal distribution or a noise normal distribution and d' represents the difference in means of these distributions. To translate the latent familiarity scores into the dichotomous decision, the model assumes there is a threshold, λ, and if the familiarity is lower than that threshold people classify the stimulus as noise while if the familiarity is higher than the threshold people will classify the stimulus as signal.

When estimating only these two parameters, the SDT model has been quite popular (a google scholar search for papers published in the last ten years with key words 'signal detection theory' and 'psychology' yielded more than 20,000 results). However, this SDT model assumes that the two distributions have equal variances. Analyses of empirical data in the field of recognition memory, however, often show that the variance of the signal distribution is larger than the variance of the noise distribution (e.g., DeCarlo, 2010; Macmillan & Creelman, 2005; Mickes et al., 2007; Starns & Ratcliff, 2014; Swets, 1986). Unfortunately, adding a third parameter σ (for the ratio of the variance of the signal to noise distribution) to the SDT model creates an identifiability problem; three parameters (i.e., d', λ, and σ) are estimated using only two data points (i.e., hit rate and false alarm rate). To estimate the extra parameter the model needs more informative data. One way of obtaining more informative data is by having participants rate the familiarity of each item on a confidence rating scale (e.g., "how confident are you that the word pre-

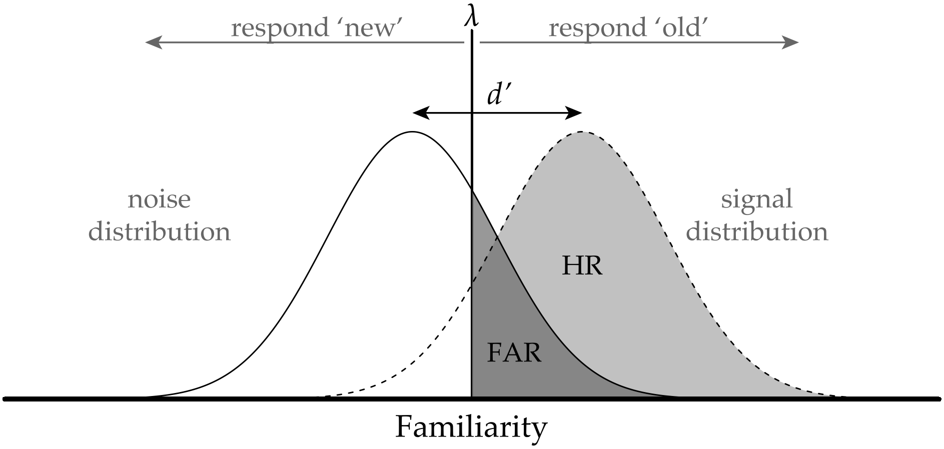


Figure 2.1: The interaction between the two parameters d' and λ lead to a certain Hit Rate (HR) and False Alarm Rate (FAR). Increasing the decision criterion leads to a lower FAR but also a lower HR, while d' stays the same.

sented was on the study list?", indicated on a Likert scale from 1–7) instead of asking for dichotomous answers ("was the word on the study list or not?"). However, with confidence rating data the number of thresholds that need to be estimated increases with the number of categories. For instance, if the SDT model is fit to data from a four-point Likert scale, this requires estimation of five parameters—d', σ, and three thresholds—but if the model were fit to data from a ten-point Likert scale, this requires estimation of eleven parameters—d', σ, and nine thresholds. The estimation of additional thresholds requires larger data sets; to estimate thresholds reliably it is important that there are a certain number of observations for each category. This in turn means that models with more categories (and therefore more thresholds that need to be estimated) require a larger number of total observations. In recognition memory, accuracy decreases with successive test trials (Criss et al., 2011), limiting the number of observations any individual participant can contribute. This problem is compounded in a typical study where multiple conditions, each requiring many observations, are under investigation simultaneously. Here we introduce a parsimonious method of estimating the thresholds by restricting the way the thresholds can be placed. This parsimony is obtained by modeling thresholds as a linear transformation of "unbiased" thresholds, which only requires two parameters for any number of thresholds. We estimate parameters in a Bayesian way, and introduce a hierarchical extension to our model that allows the estimation of group-level parameters.

The outline of this paper is as follows. First, we will briefly elaborate on Bayesian methods of parameter estimation. Next, we will introduce our model and the associated Receiver Operating Characteristics (ROC) curves. We will also show how our model leads to Bayesian estimates of detection measures while taking into account the uncertainty of the estimate. Lastly, we will introduce the hierarchical extension and apply the model to memory recognition data from Pratte et al. (2010).

2.1 Modeling the Thresholds

The key concepts in our SDT threshold model are summarized in Figure 2.2. This figure represents an example where an individual observer rated how familiar six items—three signal items and three noise items—are on a Likert scale from one to six. The model describes the process with which these data are generated. The model assumes that the observer makes internal appraisals of the familiarity of the noise items $f^{(n)}$ and the signal items $f^{(s)}$, both of which are latent and continuous. These appraisals come from the noise distribution for noise items—a normal distribution with mean $\mu^{(n)}$ and standard deviation $\sigma^{(n)}$—or from the signal distribution for signal items—a normal distribution with mean $\mu^{(s)}$ and standard deviation $\sigma^{(s)}$. For reasons of identifiability we assume that the noise distribution is a standard normal distribution; i.e., $\mu^{(n)} = 0$ and $\sigma^{(n)} = 1$. Equation 2.3 describes the formal process of this step in the model.

$$f \sim \begin{cases} \mathcal{N}(0,1) & \text{if} \quad \text{noise } (f^{(n)}), \\ \mathcal{N}(\mu^{(s)}, \sigma^{(s)}) & \quad \text{signal } (f^{(s)}). \end{cases} \tag{2.3}$$

Once observers have made an internal appraisal of the familiarity of an item, they have to translate this appraisal to the ordinal Likert scale, in this case a scale from one to six. An observer is assumed to accomplish this mapping by placing thresholds λ_c (the c represents the order of the threshold) on the latent continuous scale and comparing the internal appraisal with the thresholds resulting in the ratings $x^{(n)}$ for the noise items and $x^{(s)}$ for the signal items. As shown in Figure 2.2 the internal appraisal of the familiarity of the noise items—$f_1^{(n)}$, $f_2^{(n)}$, and $f_3^{(n)}$—leads to observed ratings $x^{(n)} = (1, 2, 5)$, and the internal appraisal of the familiarity of the signal items—$f_1^{(s)}$, $f_2^{(s)}$, and $f_3^{(s)}$—leads to observed ratings $x^{(s)} = (3, 5, 6)$.

An important property of the ordinal scale is that the differences between consecutive numbers cannot be assumed equal; on a Likert scale the distance between ‘completely agree’ and ‘agree’ can be larger than the difference between ’agree’ and ’neither agree nor disagree’. Therefore, the translation between the latent continuous appraisal to the ordinal score is relatively lax, and observers are free to use the ordinal scale in different ways. For instance,

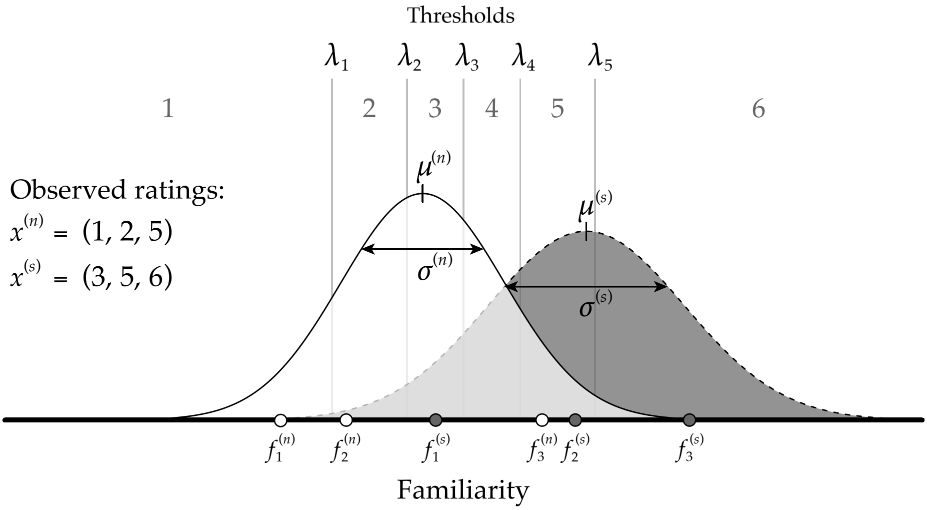


Figure 2.2: A graphical representation of the SDT threshold model for confidence ratings. Familiarity ratings are drawn from both the noise $f^{(n)}$ and the signal $f^{(s)}$ distribution. The associated confidence ratings $x^{(n)}$ and $x^{(s)}$ are generated through the thresholds λ_c.

some observers prefer to use the outer values of the scale while others prefer to use the inner values. To adjust for these individual differences, a proper model needs to be able to estimate the thresholds that are set by an observer to choose a certain answer. In previous SDT models, the number of parameters that needed to be estimated was directly related to the coarseness of the confidence scale that was used (e.g., Morey et al., 2008). Consequently, these models are not parsimonious and increase in complexity as the Likert scale becomes less coarse. In addition, the previous approaches are not easily adjusted to incorporate effect of other functional parameters (e.g., a covariate). To arrive at a more efficient way of estimating the thresholds, our model is based on a method introduced by Anders and Batchelder (2015) that uses the Linear in Log Odds function. The Linear in Log Odds function requires only two parameters to estimate a potentially large number of thresholds instead of needing a parameter per threshold (Fox & Tversky, 1995; Gonzalez & Wu, 1999). To estimate C thresholds we first assume a best-guess placement of the thresholds. First we do so on for the interval $[0, 1]$ because it is straightforward to place thresholds in an uninformative way (e.g., the intervals are of equal length). However, since the uncertainty in the SDT threshold model is expressed on the interval $[-\infty, \infty]$ we next translate the threshold placement from the $[0, 1]$ interval to the $[-\infty, \infty]$ interval.[1] Equation 2.4 shows how this translation is achieved if we were to assume that $\mu^s = 1$ and $\sigma^s = 1$. Equation 2.5 shows how these 'unbiased' thresholds are subsequently translated into the individual 'biased' thresholds using a linear transformation.

$$\gamma_c = \log\left(\frac{c/C}{1 - c/C}\right). \tag{2.4}$$

$$\lambda_c = a\gamma_c + b. \tag{2.5}$$

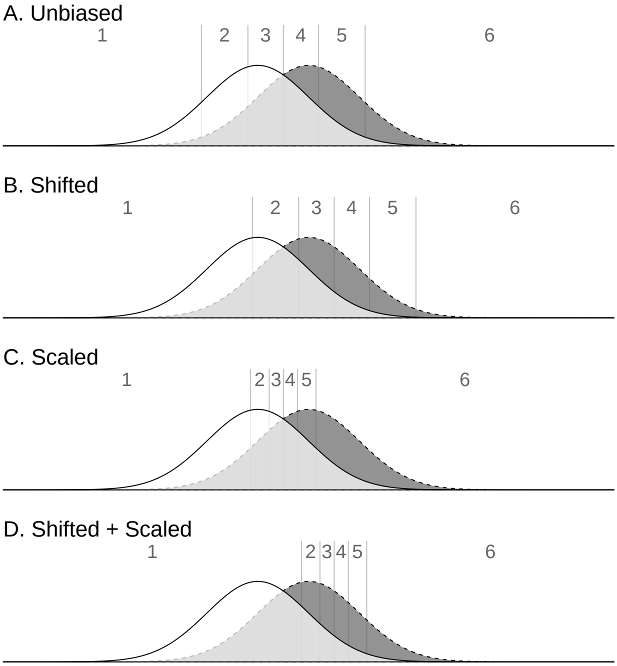


Figure 2.3: Panel A shows the position of the thresholds when an observer is 'unbiased', panel B shows the position of the thresholds when an observer prefers the lower part of the scale, panel C shows the position of the thresholds when an observer is 'unbiased' but distinguishes more between values around the center of the scale, and panel D shows the position of the thresholds when an observer prefers the lower part and distinguishes more between values where the signal distribution is high and noise distributions is low.

Here, γ_c is the unbiased threshold for each position c (e.g., γ_1 represents the first unbiased threshold). Scale parameter a allows the thresholds to be

[1] For the translation we used a logistic quantile function. Other choices, such as a Gaussian quantile function, are also possible.

distributed more closely to the center of the scale or further away from the center of the scale. Shift parameter b allows the thresholds to focus more on the left or right side of the scale and could, for example, model response bias. Figure 2.3 illustrates how these two parameters can result in different threshold placements. Compared to the unbiased thresholds in panel A, panel B shows that the thresholds have shifted to the right, and compared to the thresholds in panel B, panel C shows that the thresholds are placed closer to each other. Compared to panel C, the thresholds in panel D have shifted more to the right. This shows that two parameters can account for many different ways of threshold placement and can be extended to any number of thresholds without requiring additional parameters.

Note that the outer thresholds are always farther away from their neighboring thresholds than the inner thresholds. At first sight this may look like a major assumption of the model, but it is not. The probability of observing a certain rating is not related to the distance between thresholds, but rather to the area under the curve (i.e., the integral from one threshold to the next over either the noise or the signal distribution).

2.2 Bayesian Parameter Estimation

SDT models have been applied using both classical (Macmillan & Creelman, 2005) and Bayesian frameworks (Rouder & Lu, 2005). In this paper we adopt the Bayesian framework (Etz et al., 2016; Lee & Wagenmakers, 2013). An important goal of Bayesian statistics is to determine the posterior distribution of the parameters. This distribution expresses the uncertainty of the parameter estimates after observing the data; the more peaked this distribution the more certain the estimate. To obtain the posterior distribution of a parameter (e.g., d' or λ), the likelihood is multiplied with the prior distribution, see Equation 2.6.

$$\underbrace{p(\theta \mid \text{data}, \mathcal{M})}_{\substack{\text{posterior}\\ \text{distribution}}} \quad \overbrace{\propto}^{\text{proportional to}} \quad \underbrace{p(\theta \mid \mathcal{M})}_{\substack{\text{prior}\\ \text{distribution}}} \times \underbrace{p(\text{data} \mid \theta, \mathcal{M})}_{\text{likelihood}}. \tag{2.6}$$

In our case it is not possible to derive the posterior distribution analytically and hence we used MCMC sampling techniques (i.e. implemented in JAGS Plummer, 2003) to draw samples from the posterior distribution; with enough samples the approximation to the posterior distribution becomes arbitrarily close. As priors we used normal distributions for all unbounded parameters (mean and shift). For bounded parameters (variances and scale) we used either a gamma prior or a normal distribution truncated from 0 to ∞. Formal model definitions and prior distributions can be found in the Appendix.

To confirm the performance of the model we conducted a parameter recovery study. First, we randomly generated 100 values for $\mu^{(s)}$, $\sigma^{(s)}$, a, and b[2]. Each combination of parameters was used to generate ordinal 6-point Likert scale data (240 noise and 240 signal items), after which the SDT threshold model was fit to the data. Subsequently, we compared the parameter values used to generate the data with the means of the posterior distributions of the parameter estimates. The correlations between the data generating parameter values and the recovered parameter estimates were high ($r_{\mu^{(s)}} = 0.96$, $r_{\sigma^{(s)}} = 0.89$, $r_a = 0.99$, $r_b = 0.98$) showing that the SDT threshold model has good parameter recovery. More details on this parameter recovery study can be found in the supplemental materials at https://osf.io/v3b76/.

2.3 ROC Curve

A widely used metric to interpret parameter values of the SDT model is the Receiver Operating Characteristic (ROC) curve (Hanley & McNeil, 1982). The ROC curve displays how the hit rate and false alarm rate are affected by changes in thresholds. The translation from the SDT model parameters to the ROC curve is visualized in Figure 2.4. Each threshold in the SDT model is associated with a specific hit rate and false alarm rate. For λ_3 the hit rate is the part of the signal distribution shaded light gray, and the false alarm rate is the part of the noise distribution shaded dark gray. This associated mapping can be established for each threshold, resulting in a number of coordinates for the ROC curve. Subsequently, drawing a line through the points leads to the ROC curve.

Figure 2.5 shows three example ROC curves. In these graphs, the x-axis represents the false alarm rate and the y-axis represents the hit rate. Setting the threshold to its lowest possible value will always result in a hit or a false alarm and setting the threshold to its highest possible value will never result in a hit or a false-alarm. Therefore, the ROC curve will always go through $[0, 0]$ and $[1, 1]$. The dashed diagonal represents the hypothetical ROC curve if the signal distribution equals the noise distribution, that is, the participant is performing at chance. If the ROC curve is above the dashed diagonal this means that the participant is performing above chance, and the average strength of the signal exceeds zero.

Panel A in Figure 2.5 shows the ROC curve with near perfect detection: the hit rate reaches 1 for low values of the false alarm rate. Panel B shows a typical ROC curve when the signal and noise distribution have equal variances:

[2] The individual values for the parameters were drawn from: μ_{si}, normal distribution with mean 1 and standard deviation 0.5 truncated at $[0, 3]$, σ_{si}, normal distribution with mean 1 and standard deviation 0.5 truncated at $[1, 3]$, a, gamma distribution with shape parameter 2 and rate parameter 2, and b, normal distribution with mean 0 and standard deviation 0.5.

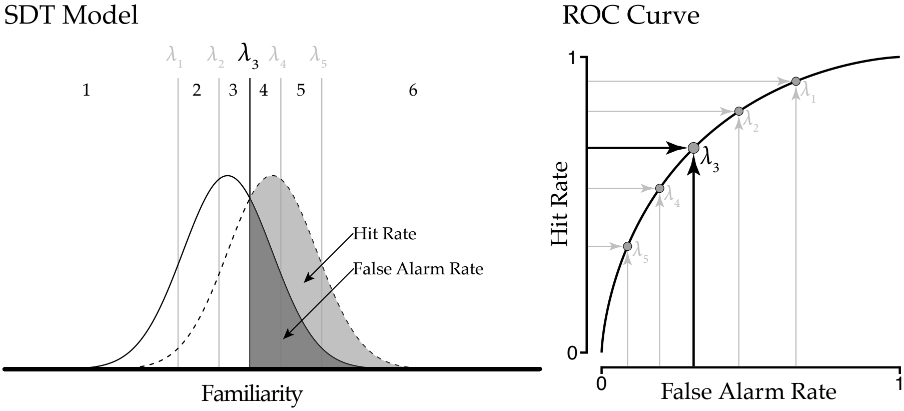


Figure 2.4: The thresholds parameters, λ_c, from the SDT model can be transformed to coordinates of the ROC curve. The hit rate and false alarm rate corresponding to each threshold can be used as coordinates for the ROC curve.

the curve is symmetrical around the minor diagonal. Panel C shows an ROC curve when the distributions do not have equal variances: the curve is not symmetrical around the minor diagonal.

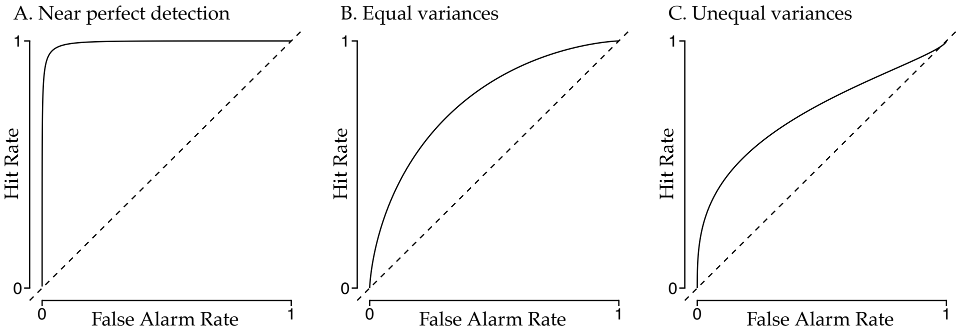


Figure 2.5: Example ROC curves. The solid line represents a theoretical ROC curve. The dashed line represents chance performance.

The mathematical relation between the SDT and ROC parameters is shown in Equation 2.7 (Marden, 1996).

$$Z_{\text{HR}} = \frac{Z_{\text{FAR}}}{\sigma^{(s)}} + \frac{\mu^{(s)}}{\sigma^{(s)}}. \tag{2.7}$$

Using this equation, the z-transformed hit rate can be calculated using the z-transformed false alarm rate, and the mean and variance of the signal distribution.[3]

2.4 Detection Measures

As we saw in the previous section, the ROC curve is able to accommodate inequality of variances. The ROC curve can easily be converted to a detection measure by calculating the Area Under the Curve (AUC Wickens, 2001); the larger the AUC, the higher the ability to detect the signal. It is clear that the AUC takes into account the inequality of variances. Also, the AUC will always be between 0.5—if detection is based purely on chance—and 1—if detection is perfect. This makes it straightforward to compare two measurements of the AUC.

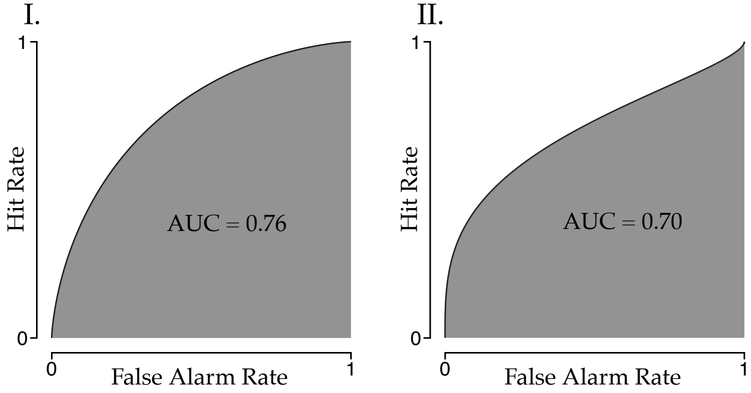


Figure 2.6: Visualization of Area Under the Curve (AUC) of an ROC curve for the two hypothetical observers. The difference in the variance of the signal distribution is expressed in the difference in the AUC.

The AUC of the ROC has the attractive property of taking into account differences in variance of the signal distribution between observers, and hence we focus on this measure. The AUC is calculated using Equation 2.8 (p. 68 Wickens, 2001), where the noise distribution is assumed to be a standard normal and Φ is the cumulative normal distribution:

$$\mathrm{AUC} = \Phi\left(\frac{\mu^{(s)}}{\sqrt{1+\sigma^{(s)2}}}\right). \tag{2.8}$$

[3] Note that z-transformed ROC functions are linear. In addition, when the equal variance assumption is met, the slope is one. When the variance of the signal distribution is larger than that of the noise distribution - as is generally found to be the case in recognition memory - the slope is less than one.

2.5 Thresholds

The most important way in which our threshold model improves upon existing confidence ratings SDT models is by estimating the thresholds in a more parsimonious way. Instead of estimating the thresholds individually, which requires one parameter per threshold, the thresholds are modeled using a linear equation. This allows for better estimates of the thresholds in the face of limited data. A consequence of this method is that the threshold placement in our model is restricted to a be linear instead of freely estimated. However, the thresholds can still be placed in a wide variety of ways. Because the threshold model takes into account that observers can set their thresholds in different ways, similar abilities in signal detection can lead to different data, underscoring the difficulties of drawing conclusions directly from the data. To illustrate this point we performed a simulation study.

To obtain plausible values for the simulation study, we first fitted the threshold SDT model to data from Pratte and Rouder (2011), who gathered confidence ratings on a memory recognition task for 97 participants (this data set is described in more detail below). Based on the estimated parameter values we chose three values of the scale parameters based on the 1^{st}, 50^{th}, and 99^{th} percentiles of the estimated values (i.e., $a_1 = 0.12$, $a_{50} = 0.84$, $a_{99} = 1.74$), and three values of the shift parameters based on the 1^{st}, 50^{th}, and 99^{th} percentiles of the estimated values (i.e., $b_1 = -0.98$, $b_{50} = 0.14$, $b_{99} = 1.10$). We used fixed values of $\mu^{(s)} = 1$ and $\sigma^{(s)} = 1$ and all possible combinations of the scale and shift parameters to simulate data from the threshold SDT model, resulting in nine different data sets. Figure 2.7 shows histograms of the simulated data. It is clear the model can describe various datasets by varying the threshold placement, even when the underlying familiarity distributions are identical.

Figure 2.7 illustrates that as the scale parameter increases (i.e., moving along the columns from left to right), more answers on the inside of the scale are given and as the shift parameter increases (i.e., moving along the rows from top to bottom), the left side of the scale is used more often. This coverage of possible outcomes makes the model nearly as flexible as having an independent parameter for each threshold while minimizing the number of parameters to estimate.

2.6 Hierarchical Extension

The threshold SDT model can be used to fit data from a single observer. However, often there is interest in the detection ability of a group of observers, which requires some sort of aggregation or pooling. One way of pooling is by aggregating the data and then fitting the model on the aggregated data.

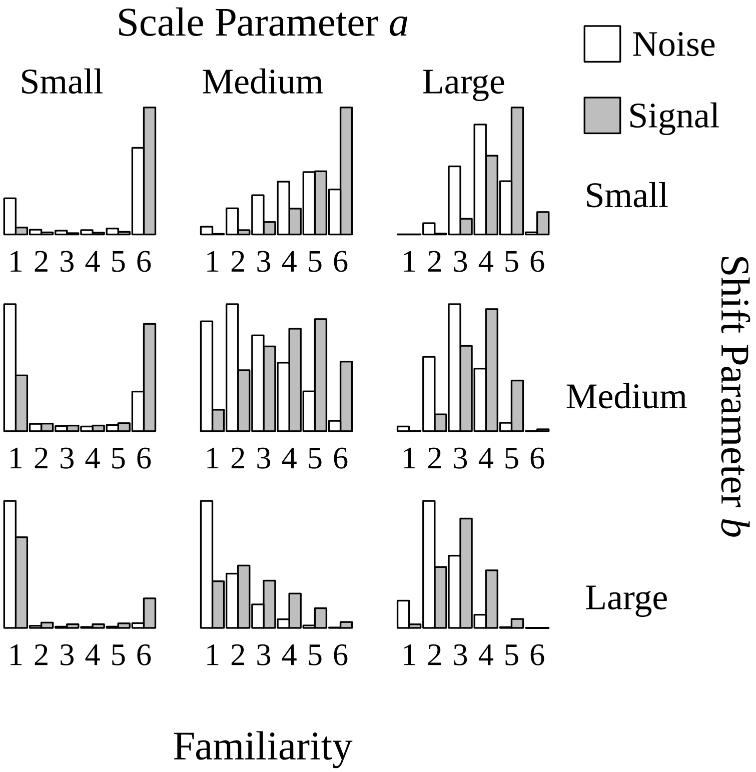


Figure 2.7: Effect of threshold parameters on familiarity judgments. Nine large datasets (N = 10,000) were simulated to visualize the range of model-implied probability distributions over familiarity judgments. The datasets were simulated with the same $\mu^{(s)}$ and $\sigma^{(s)}$, but with either a small, medium, or large scale parameter a and either a small, medium, or large shift parameter b.

Another way of pooling is by estimating the parameters for each observer individually and then take the mean or median from these parameter values. Although these methods are computationally simple, they lack a formal model that describes how the group level distribution relates to individual parameter values.

In contrast, in the Bayesian hierarchical approach, individual subject parameters are drawn from a group distribution (Gelman & Hill, 2006). Because the subjects are modeled as part of a group, the individual parameters shrink towards the group mean (Efron & Morris, 1977). The benefit of shrinkage is that the model is much more resistant to overfitting, as the group-level information makes the individual estimates less susceptible to noise fluctuations (Shiffrin et al., 2008). In the hierarchical threshold model, we introduce group distributions for the mean and variance of the signal distribution, and for the scale and shift parameters of the thresholds. The priors for unbounded parameters (mean and shift) are normal distributions whereas the priors for bounded parameters (variance and scale) are either gamma distributions or truncated normal distributions. Exact model specifications and priors are shown in the Appendix[4].

To confirm the performance of the model we conducted a parameter recovery study. The formal model definitions including prior distributions can be found in the Appendix. First, we fitted the hierarchical SDT threshold model to the data of Pratte et al. (2010) (see next section for a more elaborate explanation). We used the means of the posterior distributions for the individual level parameters $\mu^{(s)}$, $\sigma^{(s)}$, a, and b to generate plausible data. Next, we fit the model to the synthetic data and drew posterior samples from the hierarchical SDT threshold model. Subsequently, we compared the data-generating parameter values to the means of the posterior distributions for the parameter estimates. The correlation between the data-generating parameter values and the recovered parameter estimates was high ($r_{\mu^{(s)}} = 0.96$, $r_{\sigma^{(s)}} = 0.90$, $r_a = 0.99$, $r_b = 0.99$, see Figure A.6) showing that the hierarchical SDT threshold model has good parameter recovery. More details on this parameter recovery study can be found in the supplemental materials at https://osf.io/v3b76/. The next section applies the model to experimental data.

2.7 Application to Experimental Data

We fitted the hierarchical SDT threshold model to data from Pratte et al. (2010) who had gathered confidence ratings on a memory recognition task from 97 participants. Each participant studied 240 words— each word for 1,850 ms with 250 ms blank periods between two words—randomly selected from a set of 480 words. After the study phase, participants had to indicate how confident they were that a word was part of the study list on a 6-point Likert scale (using the ratings “sure new”, “believe new”, “guess new”, “guess studied”, “believe studied”, and “sure studied”) for the whole batch of 480 words. In this experiment, the words in the study list represent the signal items, while the words that were not in the study list represent the noise items.

Figure 2.8 shows the estimated median and 95% credible intervals for each parameter in the model. The dashed vertical line represents the median of the

[4] We opted not to use highly uninformative priors as the resulting prior ROC curves are implausible. See Figures A.4 and A.5 for the prior and posterior ROC curves under slightly informed and highly uninformative priors. Different priors had negligible effect on the posterior distribution, see the supplementary Figures on https://osf.io/v3b76/ for a comparison.

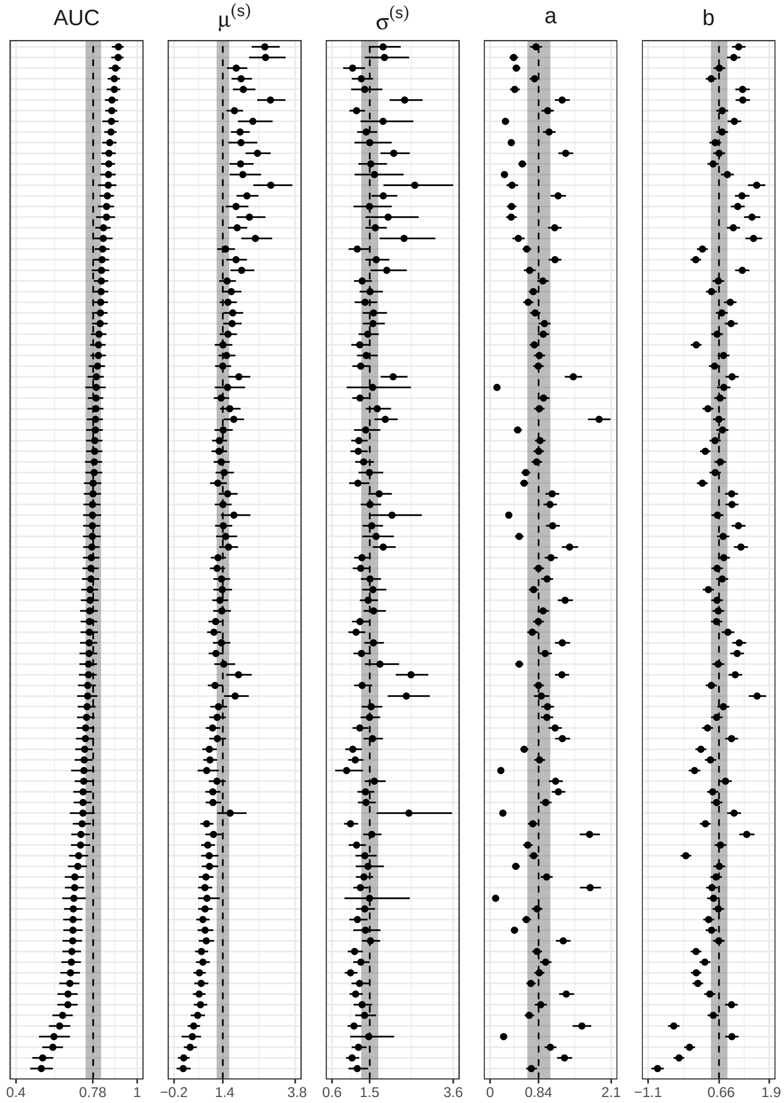


Figure 2.8: Parameter estimates for all 97 participants from Pratte et al. (2010); the dot represents the median and the line represents the 95% central credible interval. The dashed line represents the median of the group distribution and the accompanying 95% credible interval is indicated in grey.

group level estimation with the 95% credible interval shaded gray. The parameters are estimated with a good precision; in general, the credible intervals are narrow.

The model parameters can also be used to produce an ROC curve. Figure

2.9 shows the ROC curve for the group level, where the shaded area represents the uncertainty in the estimate, and the density plot shows the posterior distribution for the AUC. Note that the uncertainty in the ROC and the AUC is induced by the uncertainty in the model parameters.

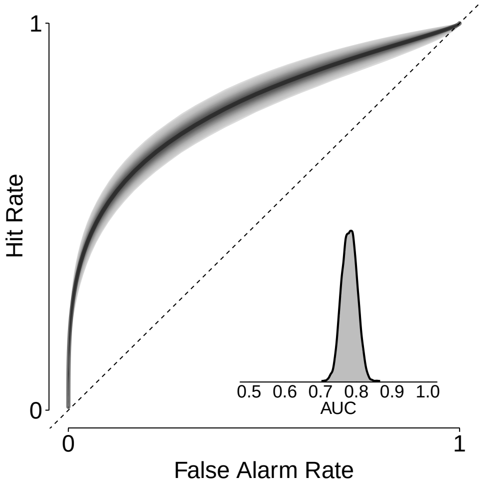


Figure 2.9: Group level ROC curve with the 95% credible interval in grey and the Area Under the Curve (AUC) with the uncertainty in the estimate expressed through the posterior distribution.

2.8 Discussion

The threshold SDT model describes how people estimate the familiarity of signal and noise items. The main contribution of the model is that it provides a parsimonious way of estimating the thresholds instead of sacrificing one parameter per threshold. We also showed how this model can be applied to experimental data. This paper presents a first effort in parsimonious threshold estimation that should be applicable to many SDT applications. It can also be used as a starting point for more complicated applications of SDT models. A straightforward empirical test of the threshold SDT model is to examine how experimental manipulations map onto the model parameters. For example, one may conduct a test of specific influence and examine the extent to which effects of changes in base-rate are absorbed by the threshold a and b parameters.

Because the threshold SDT model features only four parameters, it is relatively straightforward to add other effects, e.g. the item effects mentioned in the discussion of Pratte and Rouder (2011). For example, a researcher could hypothesize that there is a difference in response bias between two conditions, and that this difference maps onto the shift parameter. To incorporate this into the model, Equation 2.5 could be modified to include a covariate on the shift of the thresholds. Such a modification is identical to adding a predictor to a regression model. This allows for relatively easy group comparisons; in contrast, such comparisons are difficult for models that require one parameter per threshold, as multiple estimates need to be considered simultaneously.

Expanding the transformation of the thresholds into a linear model introduces the need for model comparison. To assess the relevance of a predictor one compares a model without the predictor to a model with the predictor. Within the Bayesian framework, comparing models is often done by means of Bayes factors (Jeffreys, 1961; Mulder, 2016). Although no analytical formulas exist for calculating Bayes factor for SDT models, an approximation can be obtained using numerical techniques on the obtained MCMC samples, e.g. via bridge sampling. (Gronau, Sarafoglou, et al., 2017; Meng & Wong, 1996).

In sum, the threshold SDT model provides a parsimonious and straightforward account of confidence rating data, allowing researchers to quantify not only discriminability but also confidence category thresholds. The uncertainty in the model's parameter estimates can be used to induce uncertainty in crucial SDT measures such as the area under the ROC curve.

3

Cultural Consensus Theory for the Evaluation of Patients' Mental Health Scores in Forensic Psychiatric Hospitals

In many forensic psychiatric hospitals, patients' mental health is monitored at regular intervals. Typically, clinicians score patients using a Likert scale on multiple criteria including hostility. Having an overview of patients' scores benefits staff members in at least three ways. First, the scores may help adjust treatment to the individual patient; second, the change in scores over time allows an assessment of treatment effectiveness; third, the scores may warn staff that particular patients are at high risk of turning violent, either before or after release. Practical importance notwithstanding, current practices for the analysis of mental health scores are suboptimal: evaluations from different clinicians are averaged (as if the Likert scale were linear and the clinicians identical), and patients are analyzed in isolation (as if they were independent). Uncertainty estimates of the resulting score are often ignored. Here we outline a quantitative program for the analysis of mental health scores using cultural consensus theory (CCT Anders & Batchelder, 2015). CCT models take into account the ordinal nature of the Likert scale, the individual s among clinicians, and the possible commonalities between patients. In a simulation, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and, as an alternative, against often-used machine learning toolboxes. In addition, we outline the substantive conclusions afforded by the application of the CCT model. We end with recommendations for clinical practitioners who wish to apply CCT in their own work.

This chapter is published as: van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology, 98*, 102383.

Forensic psychiatric hospitals monitor the mental health and forensic risk factors of their patients at regular intervals, typically using a method such as Routine Outcome Monitoring (de Beurs et al., 2011). A clinician, psychiatrist, or another staff member, henceforth a *rater*, scores a patient on historical, clinical, and prospective criteria. For example, a rater evaluates a patient's risk factors and behavior on a variety of criteria that relate to aggressiveness and the risk of recidivism. Such evaluations are stored so they may be used to inform future decisions. The decisions informed by these ratings can vary widely. For instance, the scores may help adjust treatment to individual patients, the change in scores over time allows for an assessment of treatment effectiveness, and the scores may warn staff that particular patients are at high risk of turning violent. Moreover, these ratings are key for a quantitative approach to monitoring and forecasting patients' behavior.

Current practices for aggregating the scores are suboptimal. Evaluations from different raters are often averaged as if they are exchangeable. For example, personal communication with the staff of a forensic psychiatric hospital suggested that clinicians are more lenient in their ratings than psychiatrists, but this information is not used to weigh their ratings. Furthermore, patients are analyzed in isolation, as if they are independent of one another. For example, consider a random sample consisting of patients with a schizophrenic disorder and patients with an addictive disorder. The patients are clearly not independent of each other; patients with the same disorder will most certainly resemble each other more. Any background information about patients, such as a patient's criminal record, is not accounted for and is only seen as static baseline information. In addition, uncertainty estimates of the resulting score are usually ignored.

Here we address these issues using Cultural Consensus Theory (CCT; Batchelder & Anders, 2012; Batchelder & Romney, 1988; Romney et al., 1986). The defining characteristic of CCT is that it aims to estimate the consensus knowledge shared by raters. Hence, CCT is a promising framework for analyzing data of forensic psychiatric hospitals, where the true state of a patient is unknown and needs to be estimated from the scores given by the raters. CCT models capture individual differences between raters and items, and pool information while accounting for these differences. However, currently available CCT models can only be applied to the data of a single patient; a limitation addressed in this paper.[1]

The focus of this paper is to outline a quantitative program for the analysis of mental health scores using CCT. First, a CCT model for ordinal data is introduced (Anders & Batchelder, 2015). Next, this model is expanded step

[1] To be precise, currently available CCT models can only describe two hierarchical structures, i.e., for data of patients and raters, patients and items, or items and raters. However, existing CCT models treat the third hierarchical structure as non-hierarchical.

by step, to allow a more sophisticated account of the data, for instance by describing multiple patients. We showcase the model in three simulation studies. First, we illustrate the benefits of this approach by analyzing two fictitious patients. Second, we show that model parameters are retrieved accurately. Third, we compare the predictive performance of the CCT model to the current practice of aggregating raw observations and against often-used machine learning toolboxes such as Random Forest (Breiman, 2001) and Boosted Regression Trees (Friedman, 2002). We highlight the substantive conclusions obtained from applying the CCT model and conclude the paper with recommendations for clinical practitioners who wish to apply CCT in their own work.

3.1 Cultural Consensus Theory and Three Extensions

The next sections introduce Cultural Consensus Theory (CCT). First, a brief introduction to CCT is given. Next, the CCT model developed in Anders and Batchelder (2015, henceforth AB) is introduced, which serves as the simplest model for a single patient. Subsequently, we generalize the model in three ways. First, the model is expanded to describe multiple patients simultaneously. Next, latent constructs are added to the model. Finally, the model is adapted to include background information on patients and raters.

3.1.1 Cultural Consensus Theory

Cultural Consensus Theory, also known as “test theory without an answer key” Batchelder and Romney, 1988, is a statistical tool that can be used to retrieve the unknown “truth” for an item by examining the consensus among the responses. For example, given a political questionnaire, there are no objectively correct answers. Instead, one could administer the questionnaire to left-oriented respondents and use CCT to find out what the consensus is among left-oriented respondents. CCT models can capture that some responders have a higher competency and will strictly answer according to the cultural consensus. Likewise, items can differ in their difficulty, i.e., the competence required to answer according to the consensus. For a political questionnaire, this implies that only extremely left-oriented respondents agree with the most left-oriented political statements. Note that competence and difficulty parameters are relative to the consensus and do not refer to absolute competence or difficulty. Instead competence captures the extent to which a rater evaluates according to the group consensus; likewise, difficulty captures how high a rater’s competence must be to be expected to answer an item according to the group consensus. In addition, CCT models can be expanded to allow for multiple consensus truths, that is, there can be multiple unknown truths that vary across subgroups of respondents (Anders & Batchelder, 2012). For a political questionnaire, the different consensuses (e.g., left, right, center, etc.) and respondents membership to these groups would be estimated from the data. The property of CCT models to estimate the consensus truth from the data is ideal for psychiatric data, where a patient's true state is unknown and a consensus from the raters is desired. CCT models can be applied to continuous data (e.g., the LTM Batchelder & Anders, 2012), categorical data (e.g., the General Condorcet model Batchelder & Romney, 1986), and ordinal data (AB). Since ratings are usually given on a Likert scale, we focus on a CCT model for ordinal data.

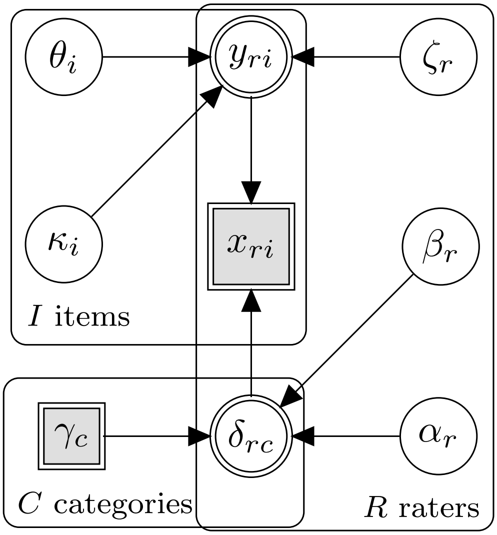
3.1.2 The Latent Truth Rater Model

As a starting point, consider the Latent Truth Rater Model (LTRM), a cultural consensus model for ordinal data introduced by AB. Figure 3.1 shows a graphical model of the LTRM and Table 3.1 provides an overview of the parameters. The LTRM captures differences among raters and items and may be viewed as the simplest model for a single patient.

The rating of rater r on item i is denoted x_{ri} and takes on discrete values from 1 through C. AB formalize the core ideas of the LTRM with 6 axioms, which are briefly repeated here. There is an unknown latent shared cultural truth among the raters, which is captured by the item location parameters θ_i (AB's axiom 1). Since raters are not perfect measurement instruments, they infer a noisy version of the cultural truth for each item, called a latent appraisal and defined as $y_i = \theta_i + \epsilon_{ri}$, where $\epsilon_{ri} \sim \text{Logistic}\left(0, {}^{\zeta_r}/_{\kappa_i}\right)$ (AB's axiom 2). The logistic density with location l and scale s is defined as

$$\text{Logistic}\left(x;\, l,\, s\right) = \frac{\exp\left(-\frac{x-l}{s}\right)}{s\left(1+\exp\left(-\frac{x-l}{s}\right)\right)^2} \quad \text{where } s > 0.$$

The scale of the logistic distribution for the latent appraisals consists of two components. Differences in item difficulty are captured by κ_i and differences in rater competence are captured by ζ_r (AB's axiom 3). The ratio of item difficulty over rater competence is the variance of the latent appraisal. For example, if an item is difficult then the variance of the latent appraisals is high, which leads to a spread-out probability distribution over observed ratings. Likewise, if the rater competence is high, then the variance of the latent appraisals is low and the probability distribution over observed ratings is concentrated. Latent appraisals y_{ri} are assumed to be conditionally independent given the latent truth θ_i, the item difficulty κ_i, and the rater competence ζ_r (i.e., their joint distribution can be factored into a product of univariate distributions that only depend on the three aforementioned parameters; AB's axiom



$$x_{ri} = \begin{cases} 1 \text{ if } y_{ri} \le \delta_{r1} \\ c \text{ if } \delta_{r,c-1} < y_{ri} \le \delta_{rc} \\ C \text{ if } y_{ri} > \delta_{r,C-1} \end{cases}$$

$$\begin{aligned} y_{ri} &= \theta_i + \epsilon_{ri} \\ \gamma_c &= \text{logit}\left(c/C\right) \\ \delta_{rc} &= \alpha_r \gamma_c + \beta_r \\ \epsilon_{ri} &\sim \text{Logistic}\left(0,\, \kappa_i/\zeta_r\right) \\ \theta_i &\sim \text{Normal}\left(\mu_\theta,\, \sigma_\theta^2\right) \\ \kappa_i &\sim \text{Gamma}\left(\mu_\kappa^2/\sigma_\kappa^2,\, \mu_\kappa/\sigma_\kappa^2\right) \\ \alpha_r &\sim \text{Gamma}\left(\mu_{\alpha_r}^2/\sigma_{\alpha_r}^2,\, \mu_{\alpha_r}/\sigma_{\alpha_r}^2\right) \\ \beta_r &\sim \text{Normal}\left(\mu_{\beta_r},\, \sigma_{\beta_r}^2\right) \\ \zeta_r &\sim \text{Gamma}\left(\mu_{\zeta_r}^2/\sigma_{\zeta_r}^2,\, \mu_{\zeta_r}/\sigma_{\zeta_r}^2\right) \end{aligned}$$

Figure 3.1: Graphical model corresponding to the LTRM; a CCT model for a single patient. x_{ri} is an observed response, y_{ri} is the underlying continuous latent appraisal, θ_i is the underlying latent appraisal, and ϵ_{ri} is the appraisal error. Furthermore, κ_i and ζ_r capture the item difficulty and rater competence respectively. The unbiased thresholds are denoted γ_c, the scale and shift parameters are α_r and β_r respectively. The transformed thresholds are denoted δ_{rc}. The group-level means and standard deviations are denoted μ and σ respectively. The priors on the group-level parameters are omitted. Gamma distributions are parametrized with shape and scale so that the group-level parameters correspond to the mean and standard deviation of the distribution.

4). So far, the axioms describe a continuous latent process that underlies each observation. To translate these continuous latent appraisals to categorical responses, it is assumed that there exist $C-1$ ordered thresholds δ_{rc}, such that each x_{ri} is generated deterministically in the following way (AB's axiom 5):

$$x_{ri} = \begin{cases} 1 & \text{if } y_{ri} \le \delta_{r1} \\ c & \text{if } \delta_{r,c-1} < y_{ri} \le \delta_{rc} \\ C & \text{if } y_{ri} > \delta_{r,C-1} \end{cases}$$

where $c = 1, \ldots, C$. The appraisal y_{ri} is latent and thus we consider the probability that an appraisal falls between two thresholds to obtain the probability of an observed score. This makes the generating process of x_{ri} probabilistic and described by an ordered logistic distribution[2], which gives:

$$P(x_{ri} \mid y_{ri}, \delta_r) = \begin{cases} 1 - F\left(y_{ri} - \delta_{r1}\right) & \text{if } x_{ri} = 1, \\ F\left(y_{ri} - \delta_{r,c-1}\right) - F\left(y_{ri} - \delta_{rc}\right) & \text{if } 1 < x_{ri} < C, \\ F\left(y_{ri} - \delta_{r,C-1}\right) & \text{if } x_{ri} = C. \end{cases}$$

where $F\left(x\right) = \left(1 + \mathrm{e}^{-x}\right)^{-1}$, the cumulative distribution function of the standard logistic distribution. The thresholds δ_{rc} accommodate the response biases of the raters. AB do so by estimating $C - 1$ ordered thresholds γ and defining $\delta_{rc} = \alpha_r\gamma_c + \beta_r$ (AB's axiom 6). This translation of thresholds is called the Linear in Log Odds function and is a useful tool for capturing bias in probability estimation (Anders & Batchelder, 2015; Fox & Tversky, 1995; Gonzalez & Wu, 1999). Specifically, the scale parameter concentrates the thresholds closer together or farther apart, and thus can yield a flat or peaked probability distribution, respectively. The shift parameter β moves all thresholds up and down relative to the item location and thus captures the fact that some raters give higher overall ratings than others.

Figure 3.2 provides an intuition for how the ordered logistic distribution can model different outcomes by varying only the rater parameters. The latent appraisal y is fixed to 0, the thresholds γ are equal to $\operatorname{logit}\left(c/C\right)$ such that $P(x_{ri} \mid y = 0,\ \gamma,\ \alpha_r = 1,\ \beta_r = 0)$ is uniform, and the scale α_r and shift β_r vary. In the left panel, there is no response bias, $\alpha_r = 1$ and $\beta_r = 0$, which yields a uniform distribution over the predicted Likert scores. In the right panel, an increase in response scale and shift, $\beta_r = .5$ and $\alpha_r = 2$, concentrates the predicted Likert scores around 2 and 3.

The LTRM is a complex model and unfortunately suffers from identification issues, as AB already pointed out. For example, multiplying the rater competences ζ and the item difficulties κ by a constant c yields an identical variance for the appraisal distribution since ${}^{c\zeta}/_{c\kappa} = {}^{\zeta}/_{\kappa}$. Such identification problems are avoided by restricting the mean of the respective parameters to 1 (as suggested in Appendix C in AB). Another identification problem originates from estimating the thresholds individually. The number of thresholds, $C - 1$, increases with the number of response options. This introduces a large number of parameters that can be difficult to estimate, in particular when some response options are not observed (i.e., when there are ceiling or floor effects). In addition, the model is only identified if the sum of thresholds is zero ($\sum_{c=1}^{C} \gamma_c = 0$; otherwise adding a constant to θ_i and δ_c yields an identical likelihood). Rather than modeling each threshold individually, we describe the thresholds using only two parameters per rater. Specifically, we model

[2] The choice for an ordered logistic distribution is arbitrary and an ordered probit distribution could also be used, as was done by AB. We use a logistic distribution rather than a normal distribution because its cumulative distribution function has an analytic expression.

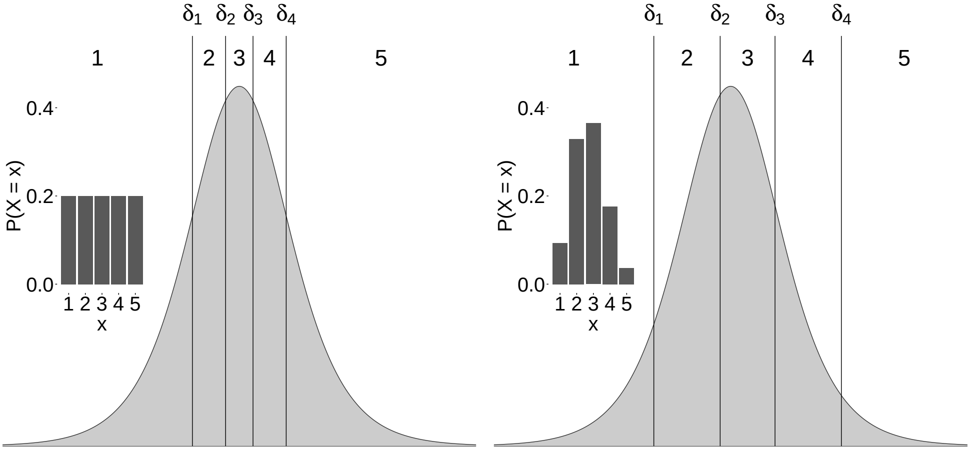


Figure 3.2: The ordered logistic distribution relates latent appraisals y_{ri} to response categories $\gamma = 1, \ldots, 5$ via category thresholds $\delta_1, \ldots, \delta_4$. The implied probability distribution over response categories is shown inside each panel. In the left panel, there is no response bias, $\alpha_r = 1$, $\beta_r = 0$. As a consequence, the distribution over the predicted Likert scores is uniform. In the right panel, the thresholds are shifted right, $\beta_r = 0.5$, and the scale increased slightly, $\alpha_r = 2$, such that the distribution over predicted Likert scores is peaked on outcomes 2 and 3. In both panels, the item location parameter θ_i is 0.

the thresholds as deviances from an initial guess, $\gamma_c = \text{logit}\,(c/C)$. This yields a set of thresholds such that if the latent appraisal is 0 then $P(x_{ri})$ is uniform. Response biases are incorporated in the same manner: $\delta_{rc} = \alpha_r \text{logit}\,(c/C) + \beta_r$. This simplification can still capture a wide variety of data sets (Selker et al., 2019).

3.1.3 Three Extensions

The LTRM as described above has many desirable properties; for instance, it captures individual differences among both raters and items. However, many properties of psychiatric data are not captured by the model. Three extensions generalize the LTRM to improve its capacity to describe the data at hand.

Extension I: Multiple Patients

The first extension allows the model to describe multiple patients instead of a single patient. Since even patients with the same disorder can have different ratings for the same item, the latent truth for an item varies across patients to

Table 3.1: Overview of the parameters in the LTRM. The first column indicates the parameter, the second the parameter bounds, and the third provides the definition or prior distribution of that parameter. The last column provides a brief description of the parameter.

Parameter	Domain	Definition/ prior	Meaning
y_{ri}	$\mathbb{R}$	$\theta_i + \epsilon_{ri}$	Appraisal of rater r on item i.
γ_c	$[0,\, 1]$	$\text{logit}\,(c/C)$	Unbiased thresholds for outcome c.
δ_{rc}	$\mathbb{R}$	$\alpha_r\gamma_c + \beta_r$	Transformed thresholds for rater r on outcome c.
ϵ_{ri}	$\mathbb{R}$	Logistic $(0,\, \kappa_i/\zeta_r)$	Residual of appraisal.
θ_i	$\mathbb{R}$	Normal $\left(\mu_\theta,\, \sigma_\theta^2\right)$	Location of item i.
κ_i	$\mathbb{R}^+$	Gamma $(\mu_\kappa^2/\sigma_\kappa^2,\, \mu_\kappa/\sigma_\kappa^2)$	Difficulty of item i.
α_r	$\mathbb{R}^+$	Gamma $(\mu_{\alpha_r}^2/\sigma_{\alpha_r}^2,\, \mu_{\alpha_r}/\sigma_{\alpha_r}^2)$	Scale-bias of rater r.
β_r	$\mathbb{R}$	Normal $\left(\mu_{\beta_r},\, \sigma_{\beta_r}^2\right)$	Shift-bias of rater r.
ζ_r	$\mathbb{R}^+$	Gamma $\left(\mu_{\zeta_r}^2/\sigma_{\zeta_r}^2,\, \mu_{\zeta_r}/\sigma_{\zeta_r}^2\right)$	Competence of rater r.

reflect this. Likewise, it can be more difficult for raters to answer specific items according to the consensus, but only for some patients. To describe parameters that vary across patients we introduce the subscript p for patient. Both these changes can be achieved by allowing the item truth θ_{ip} and item difficulty κ_{ip} to vary across patients, so that the latent appraisal y_{rip} varies across patients. Note that item difficulty is no longer specific to items, but also captures the interaction between patients and items. Modeling this interaction is useful when, for example, a patient barely cooperates with a question about his or her feelings; as a result, it is hard to score this item according to the consensus, but only for this patient. As in Figure 3.1, we assume that the patient parameters are drawn from a group-level distribution with unknown mean and variance, for instance, the item difficulty could follow a gamma distribution with unknown mean and variance (i.e., $\kappa_{ip} \sim \text{Gamma}\,(\mu_\kappa^2/\sigma_\kappa^2,\, \mu_\kappa/\sigma_\kappa^2)$).

Extension II: Latent Constructs

Often, we are not just interested in the latent truth of a single item, but also in a construct that is measured by multiple items. For instance, the latent construct aggressiveness could be measured with multiple items. To allow the model to measure constructs, we introduce a latent variable η_{pl} that represents the score of patient p on latent variable l. Items can load on different latent variables, which introduces a factor model over the items. The relation between the latent construct l and the item consensus i is given by the factor loading λ_{il}, such that $\theta_{ip} \sim \text{Normal}\left(\lambda_{il}\eta_{pl},\, \sigma^2_{\eta_p}\right)$. The measurement model, i.e., which items load on what latent construct, is assumed to be known.

As prior distribution on the latent constructs η_{pl} we used a normal distribution with mean 0 and variance 1, which reflects that the mean and variance of a latent variable are typically unidentified. In addition, simulations showed that the estimated regressions weights and the estimated patients' scores on the latent constructs exhibited label switching. For example, multiplying both the latent constructs η and the factor loadings λ by -1 yields the same distribution over the item truths. To avoid label switching, we restricted the factor loadings to be positive. Since we assume the factor structure to be approximately known, items that will have a negative loading on the latent construct can be reverse-scored. Here, approximately implies that if an item loads on a scale, we know whether it correlates positively or negatively with the scale although the magnitude is unknown.

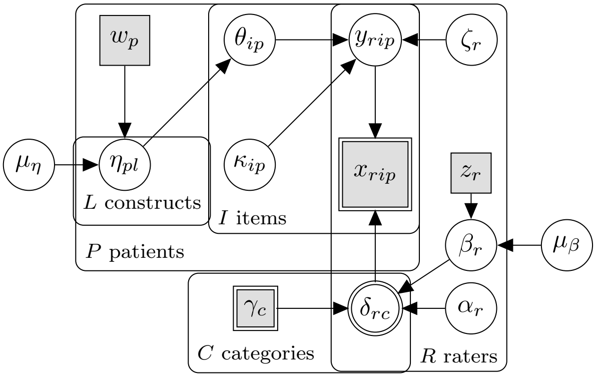
Extension III: Patient and Rater Information

The third extension adds background information about raters and patients to the LTRM. This helps the model to capture that, for instance, patients with a pedophilic disorder are typically less aggressive than murderers. Discrete patient characteristics, such as criminal record, and rater characteristics are captured by introducing separate parameters of the group-level distributions for each level of the discrete characteristic. For example, the mean of the group-level distribution of the aggressiveness scale is estimated separately for murderers and patients with a pedophilic disorder. More formally, background information is represented by a categorical indicator w_p that takes on values 1 through D for each patient p. The group-level distribution for factor scores then becomes $\eta_{pl} \sim \text{Normal}\left(\mu_{w_p l},\, \sigma_{w_p l}\right)$.

Rater characteristics are denoted z_r and are incorporated in similar manner. Rater characteristics influence the group-level distributions of rater-specific parameters, which yields $\beta_r \sim \text{Normal}\left(\mu_{z_r},\, \sigma_{z_r}\right)$. For instance, this could capture that clinicians give more lenient ratings than psychiatrists. Similarly, the group-level distribution of α_r could also be modeled as a function of rater characteristics. However, we did not include this in the model as there was no empirical observation that implies the scale parameters differ across groups of raters.

In the simulation studies, we restrict the analysis to discrete background information. However, continuous background information could also be used. Consider for instance the time a patient is committed to a psychiatric hospital, Time_p. This information can be added as a regression on the mean of the group-level distribution. Thus, $\eta_{pl} \sim \text{Normal}\left(\mu_{w_{pl}} + \nu\,\text{Time}_p,\, \sigma_{w_{pl}}\right)$, where ν is the regression coefficient from the time a patient is committed Time_p on the mean of the group-level distribution.

It is important to consider that the influence of background variables can differ across latent constructs. For instance, the effect of a patient's crime varies across latent constructs, allowing the model to capture that patients with a pedophilic disorder and murderers differ in aggression, but not on depression. This is accomplished by estimating the effect of a patient's crime separately for each latent construct.



$$\begin{aligned}
\lambda_{il} &\sim \text{Normal}^{+}(0,\, 10)\\
\theta_{ip} &\sim \text{Normal}(\lambda_{il}\eta_{pl},\, 1)\\
\eta_{pl} &\sim \text{Normal}\left(\mu_{w_{pl}},\, \sigma_p^2\right)\\
\beta_r &\sim \text{Normal}\left(\mu_{z_r\beta},\, \sigma_{\beta_r}^2\right)\\
\mu_\beta &\sim \text{Normal}(0,\, 10)\\
\mu_\eta &\sim \text{Normal}(0,\, 10)
\end{aligned}$$

Figure 3.3: Graphical model corresponding to the CCT model for multiple patients. The data x_{rip}, latent appraisals y_{ri}, latent truths θ_{ip}, and item difficulty κ_{ip} now vary across patients, as indicated by the subscript p. Furthermore, for raters, competence, scale and shift are captured by ζ_r, α_r, and β_r respectively. The unbiased and biased thresholds are denoted γ_c and δ_{rc}. Rater and patient covariates are represented by z_r and w_{pl}, whereas their effects are captured by the vectors μ_β and μ_η respectively. The prior distributions are shown on the right for modified parameters. Priors not shown can be found in Figure 3.1. The prior distributions for the extended LTRM were chosen to be weakly informative.

Figure 3.3 graphically summarizes the extended LTRM and Table 3.2 provides an overview of the parameters. The extended LTRM first separates the rater-specific influences from the data x_{rip}, hereby accounting for different groups of raters. This results in a latent consensus for each item and patient θ_{ip}. This consensus is subsequently used as an indicator for a latent construct for all patients and constructs η_{pl}. The relation between the latent construct and the items is given by the factor loadings λ_{il}, such that $\theta_{ip} \sim \text{Normal}(\lambda_{il}\eta_{pl},\, 1)$. The factor scores also incorporate patient-specific background information, such as the crime a patient committed.

Table 3.2: Overview of the parameters in the extended LTRM. The first column indicates the parameter, the second the parameter bounds, and the third provides the definition or prior distribution of that parameter. The last column provides a brief description of the parameter.

Parameter	Domain	Definition/ prior	Meaning
y_{rip}	$\mathbb{R}$	$\theta_{ip} + \epsilon_{ri}$	Appraisal of rater r on item i and patient p.
γ_c	$[0,\ 1]$	$\text{logit}\left({}^{c}/{}_{C}\right)$	Unbiased thresholds for outcome c.
δ_{rc}	$\mathbb{R}$	$\alpha_r \gamma_c + \beta_r$	Transformed thresholds for rater r on outcome c.
ϵ_{rip}	$\mathbb{R}$	Logistic $\left(0,\ {}^{\kappa_{ip}}/{}_{\zeta_r}\right)$	Residual of appraisal.
θ_{ip}	$\mathbb{R}$	Normal $\left(\lambda_{il}\eta_{pl},\ 1\right)$	Location of item i for patient p.
κ_{ip}	$\mathbb{R}^+$	Gamma $\left({}^{\mu_\kappa^2}/{}_{\sigma_\kappa^2},\ {}^{\mu_\kappa}/{}_{\sigma_\kappa^2}\right)$	Difficulty of item i for patient p.
α_r	$\mathbb{R}^+$	Gamma $\left({}^{\mu_{\alpha_r}^2}/{}_{\sigma_{\alpha_r}^2},\ {}^{\mu_{\alpha_r}}/{}_{\sigma_{\alpha_r}^2}\right)$	Scale-bias of rater r.
β_r	$\mathbb{R}$	Normal $\left(\mu_{z_r\beta},\ \sigma_{\beta_r}^2\right)$	Shift-bias of rater r.
ζ_r	$\mathbb{R}^+$	Gamma $\left({}^{\mu_{\zeta_r}^2}/{}_{\sigma_{\zeta_r}^2},\ {}^{\mu_{\zeta_r}}/{}_{\sigma_{\zeta_r}^2}\right)$	Competence of rater r.
λ_{il}	$\mathbb{R}^+$	Normal$^+$ $(0,\ 10)$	loading of θ_{ip} on factor η_{pl}.
η_{pl}	$\mathbb{R}$	Normal $\left(\mu_{w_pl},\ \sigma_p^2\right)$	latent construct underlying θ_{ip}.
μ_β	$\mathbb{R}$	Normal $(0,\ 10)$	Rater group effects.
μ_η	$\mathbb{R}$	Normal $(0,\ 10)$	Latent construct group effects.
w_p	$\{0, 1, \ldots\}$	Data	Indicator variable that groups patients.
z_r	$\{0, 1, \ldots\}$	Data	Indicator variable that groups raters.

3.2 Implementation

The next sections illustrate the LTRM in a variety of scenarios. First, we demonstrate the benefit of the LTRM over the raw means in an example analysis of two fictitious patients. Second, we demonstrate that the parameters of the LTRM can be accurately recovered. Last, we compare the predictive performance of the LTRM to the unweighted mean of the observations and two machine learning toolboxes.

We estimate the parameters of the LTRM and the extended LTRM using a Bayesian approach. Therefore, we are interested in the posterior distributions of the model parameters. All models were written in Stan and approximated the posterior distributions with variational inference (Carpenter et al., 2017). We opted to use variational inference over traditional Markov chain Monte Carlo because it was computationally fast while providing similar results in terms of parameter retrieval and model predictions. All data were simulated using R R Core Team, 2022 and Stan models were run using the R package `RStan` (Stan Development Team, 2019). R files and Stan models are available in the online appendix at https://osf.io/jkv38/.

3.3 Example Analysis

Here we showcase the benefits of a CCT analysis by examining results for two patients that are part of a sample of 50 fictitious patients. This example demonstrates how misleading the sample mean can be. We simulated a data set of 50 patients, 10 raters, 20 items, and 5 answer categories. The items loaded on 3 latent constructs, further referred to as aggressiveness, anxiety, and depression. A patient-specific covariate consisting of 5 categories was added to mimic the effect of a patient's criminal offense. Similarly, two categories were of raters (e.g., clinicians and psychiatrists) were simulated. Next, we selected two patients whose differences in observed means were small relative to their differences in posterior means on the latent constructs. The means for items of each construct are shown in Table 3.3. The aggregates of the raw scores suggests that these two patients might differ in aggressiveness and depression but not in anxiety. However, after fitting the extended LTRM to the data it becomes apparent that there is more to the data than what is

Table 3.3: Raw means of the observed ratings for the two patients with similar mean responses. The standard errors of the means are shown in parentheses. The means and standard errors are computed for each scale.

	Construct		
	Aggressiveness	Anxiety	Depression
Patient 1	3.86 *(0.14)*	3.04 *(0.19)*	3.65 *(0.18)*
Patient 2	3.29 *(0.17)*	3.00 *(0.18)*	2.93 *(0.21)*

shown by these averages. Using the extended LTRM, we can visualize the posterior distributions of the latent constructs for both patients, shown in Figure 3.4. The posterior distributions tell a different story than Table 3.3.

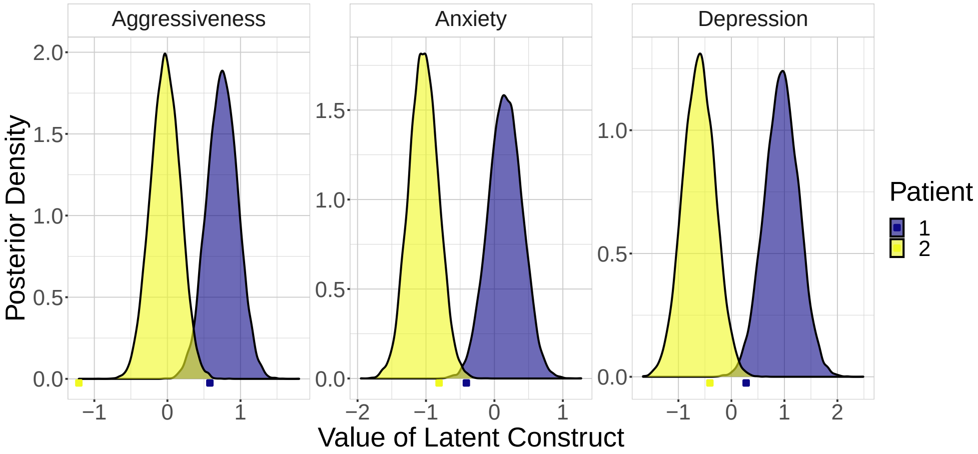


Figure 3.4: Approximate posterior densities for the η_{pl} of two patients with similar response patterns. The panels show different latent constructs. The posterior distributions suggest these patients differ on all three latent constructs, unlike what the raw means in Table 3.3 would suggest. This demonstrates that more information can be obtained from the ratings than what may be obvious from the raw scores. The squares underneath the density indicate the true values used to simulate the data.

Remarkably, for Anxiety where the raw means are approximately equal, the posterior distributions differ. This difference can be quantified by computing the posterior probability that patient 1 has a larger value on a latent trait than patient 2. This probability is approximated by counting how often the posterior samples of a latent construct are larger for patient 1 than for patient 2. For all three constructs, the probability that patient 1 has a higher score is larger than 0.99 (Figure B.1 visualizes these probabilities).

Altogether, this example shows that there is more information in the data than what the averages convey. Examining the parameters of the data generating model more closely reveals two reasons for this discrepancy. The first reason is that the item difficulty parameter κ differed among the patients for the anxiety items (the average item difficulty for anxiety was 1.42 for patient 1 and 0.88 for patient 2). The second reason is that the fictitious patients differed in background information, that is, they committed different crimes. This means that the population level distributions for the latent constructs differ for these patients.

In this example, all raters rated both patients. In practice, the ratings of different patients are likely given by different raters, which introduces a third source of bias. The discrepancy between the sample mean and posterior mean is shown for all patients in Figure B.2, which further emphasizes that the sample mean is an inadequate description of the patients' scores.

Naturally, the sample mean need not always perform this poorly. The more the data from different raters, items, and patients are exchangeable, the closer the predictions of the LTRM will be to that of the sample mean.

3.3.1 Parameter Retrieval

A key step in developing a model is to assess if the model parameters can be retrieved accurately. For this purpose, we simulated data as in the previous example; the simulated data set consisted of 50 patients, 10 raters, 20 items, and 5 answer categories. The items loaded on 3 different latent constructs. A patient-specific covariate, consisting of 5 categories was added to mimic the effect of a patient's criminal offense. Similarly, two different categories of raters were assumed. These simulation settings resemble data sets often obtained in clinical practice (e.g., Kamphuis et al., 2014). Figure 3.5 displays the true values against the posterior means for each parameter. Details and code to replicate the simulation can be found at https://osf.io/jkv38/.

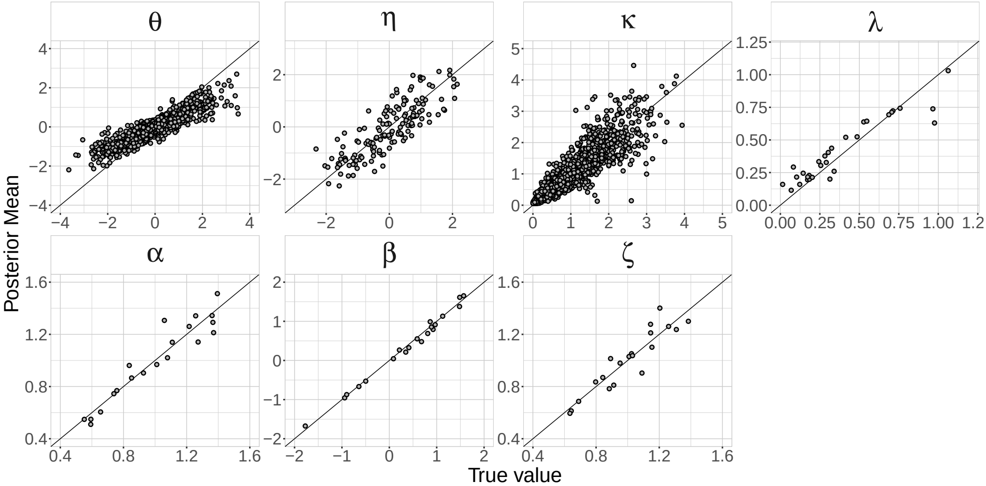


Figure 3.5: True value used for data simulation (x-axis) and posterior mean of that parameter (y-axis), for all parameters of the LTRM. Above each panel is indicated which parameter is shown.

All parameters are retrieved adequately. An exception is the item difficulty κ whose estimates appear more variable as the true item difficulty increases. The spread in posterior means of the item difficulty is similar to that in Figure 6 in AB. The item truths θ seem underestimated as their absolute magnitude increases. The hierarchical structure of the extended LTRM likely shrinks the item truths towards the mean. Typically, there is more shrinkage if the values of the parameter are larger, as is the case here. The bias in the item truths does not appear to influence the retrieval of any other parameters, for example, the latent construct scores η are retrieved accurately.

Although it is good to know when the parameters of the extended LTRM can be recovered, it may be more useful to know when the data are not sufficiently informative to apply the extended LTRM. This is likely the case when there are few items and raters. Exact numbers, however, may vary depending on the specific situation at hand. For most purposes, it is straightforward to adjust the number of raters, items, and patients, and then repeat the simulation. As an exercise, we also recovered the parameters for AB's LTRM. Code for the simulation is available in the online appendix and parameter recovery is shown in Figure B.3.

3.3.2 Predictive Performance

Here, we compare the predictive performance of the LTRM to that of the sample mode, the sample median, the sample mean rounded towards the nearest integer, and, as a more informative comparison, to Random Forest and Boosted Regression Trees (Boosting). The sample mode is the most often observed outcome (since the data are discrete). Random Forest and Boosting analyses were done using the R packages `ranger` and `gbm` respectively (Greenwell et al., 2019; Wright & Ziegler, 2017). We used the default settings for the hyperparameters in both R packages. Each method made predictions on the level of the raw data, that is, the observed ratings. Performance is assessed by quantifying the distance between the predicted ratings and the simulated true ratings.

Here we briefly introduce Random Forest (Breiman, 2001) and gradient boosted regression trees (Friedman, 2001). Random Forest and Boosting are tree-based machine learning methods that learn from a training data set in order to predict out-of-sample observations. Both methods can be used across a wide range of applications. The methods make no parametric assumptions and their predictions tend to generalize extremely well to new observations.[3] However, both Random Forest and Boosting also have downsides. Both models are so-called black boxes, that is, their parameters are statistically unidentified and do not have a meaningful interpretation. So although their predictions are often on point, they cannot answer the *how* or the *why* of the phenomena

[3] On *kaggle*, an online platform for machine learning competitions, Random Forest and Boosting are among the most successful machine learning techniques, see https://www.kaggle.com/bigfatdata/what-algorithms-are-most-successful-on-kaggle.

they predict. Furthermore, these models cannot be simulated from and they do not provide uncertainty estimates.

We simulated two data sets each consisting of 20 raters, 30 items, 50 patients, and thus in total 30,000 observations. Patients were scored on a 5-point Likert scale. The first data set represented a dense design, where all raters scored all patients. The second data set represented a sparse design, where each rater scored 10 patients, which mimics the practically plausible situation where ratings of different patients are given by different raters. Raters were pseudo-randomly assigned to patients so that the number of obtained scores was about equal for all patients. To simulate a sparse data set, we first simulated a dense data set and subsequently removed a score if the rater did not rate a patient. This remaining sparse data set consisted of 6,000 observations. Next, both data sets were split into a training set (80%) and a test set (20%). The performance of the six methods was evaluated by training the models on the training set and using the trained model to predict the outcomes for a test set. For the LTRM and the extended LTRM, we used the mode of the posterior predictive distribution as a point-prediction.[4] Predictions for Random Forest and Boosting were obtained by taking the majority vote of the trained classification trees.[5] For the observed sample mean, median, and mode we used all observations for the same rater, item, or patient.[6]

We quantified predictive performance by computing the confusion matrix between observations in the test set and predicted values; a contingency table with correct predictions on the diagonal. Prediction accuracy is defined as the proportion of correct predictions.

Given that the data were generated by the Extended LTRM, it comes as no surprise that it predicts more accurately than the other methods. However, even though data generated from the Extended LTRM is likely a gross simplification of reality, the results show that black-box machine learning methods perform somewhat adequately. This is somewhat surprising because the data at hand are ill-suited for black-box machine learning methods, as these have

[4] In this particular example, model predictions could also be interpreted as imputing missing values. If these are regarded as missing observations rather than predictions, they should be modeled as unknown discrete parameters of the model (Ch. 8; Gelman et al., 2014). That way, uncertainty about these missing observations is propagated into the parameters. Although we did not sample the missing observations from the joint posterior distribution, the code in the online appendix does show how to do this.

[5] In random Forest and Boosting, a large number of classification trees are fit to (subsets) of the data. To make a prediction, each tree makes a prediction and the most frequently predicted outcome is the final prediction.

[6] Predictions for the mode, median, and mean are obtained in the following manner. Let a negative subscript refer to all observations except that particular one, e.g., $x_{-r,ip}$ refers to $x_{1,ip}, x_{2,ip}, \dots, x_{r-1,ip}, x_{r+1,ip}, \dots, x_{R,ip}$; all observations for item i and patient p but not observation rip. Then predictions for the mode, median, or mean are obtained by taking respectively the mode, median, or mean of $x_{-r,ip}$, $x_{r,-i,p}$, and $x_{ri,-p}$.

Table 3.4: Prediction accuracy for the Extended LTRM, the LTRM, Random Forest, Boosting, the sample mean, the sample median, and the sample mode. The LTRM outperforms all other methods, but Random Forest and Boosting perform worse than the sample mode. Since the data are simulated the choices for the simulation settings are somewhat arbitrary, and different settings could yield a very accurate or very inaccurate predictive performance (e.g., by adjusting item difficulty and rater competence). Therefore, the absolute prediction error cannot be interpreted and only a relative comparison should be made. Since there were 5 possible outcomes, an accuracy of 0.2 corresponds to chance performance.

Method	Dense	Sparse
Extended-LTRM	0.52	0.41
LTRM	0.46	0.34
Sample Mode	0.43	0.33
Random Forest	0.42	0.36
Boosting	0.41	0.35
Sample Median	0.33	0.32
Sample Mean	0.23	0.27

difficulty capturing the hierarchical structure of the data which contains most of the information (but see Hajjem et al., 2014). Instead, if a lot of background information about patients and raters is available, this could likely improve their performance. However, machine learning methods do not provide interpretable models, which may be undesirable in practice because it makes it difficult to substantiate decisions.

3.4 Discussion

In this paper, we extended the Cultural Consensus model developed by Anders and Batchelder (2015) to apply to mental health scores of patients in forensic psychiatric hospitals. The original model was suited for data from a single patient and we extended this to multiple patients, latent constructs, and patient-and rater-specific covariates. The benefit of this approach is that we can obtain estimates for, for example, a patient's aggressiveness while accounting for rater bias, item-specific measurement error, and the nature of a patient's previous criminal offense. We have shown in a simulation that the parameters of the extended LTRM can be retrieved accurately.

Although the LTRM provided better predictions than black-box machine learning approaches, this is likely because the data were simulated from the

LTRM. In practice, it might be advantageous to combine the results from the LTRM with a machine learning method, as this may improve prediction accuracy. For example, augmenting a Random Forest model with features based on psychological theories resulted in a model with better predictions of human decisions than naive machine learning models and models based on psychological theories alone (Plonsky et al., 2017, 2019).However, machine learning approaches, despite their predictive power, may result in uninterpretable models which may be undesirable in psychiatric practice where decisions need to be motivated and possibly defended (e.g., when determining whether a treatment is effective or when deciding if a patient should be released). In addition, the LTRM provides richer information. For example, clinicians or psychiatrists may want to know if they rate very leniently or not. On the other hand, management might be interested in what covariates determine, for instance, the aggressiveness of patients.

Ideally, patients are monitored over some time and data from multiple measurement occasions is obtained and analyzed using the extended LTRM. Rather than applying the LTRM repeatedly to data from individual measurement occasions, all observations should be analyzed simultaneously. That way, a patient's progress may be monitored over time and predictions for the future time points could be obtained along with uncertainty estimates. To extend the LTRM to incorporate time-varying components is conceptually straightforward, but the exact properties of the time-varying components should depend on the data at hand. For example, one can imagine that the factor scores of a patient vary over time as described by a dynamic factor model (Forni et al., 2000; Molenaar, 1985). However, when patients are rated only rarely – say every six months – then the application of a sophisticated time series model is not feasible. Instead, simply estimating the difference between consecutive time points with an intercept may suffice. For these reasons, we did not explore a time series extension of the LTRM.

3.4.1 Limitations

In the LTRM, we assumed that the factor structure is known. In practice, however, this need not be the case. Estimating the factor structure from the data is possible, although such an endeavor shifts the focus of the LTRM to model selection rather than assessing the progress of patients. Furthermore, we ensured that the factor structure is identified by fixing all loadings to be positive. Strictly speaking, this restriction is stronger than needed to ensure that the model is identified. An alternative way is to fit the model without constraints and afterward relabel such that a factor solution that corresponds to one posterior mode is obtained (e.g., Erosheva & Curtis, 2017). Another more flexible approach is to view the latent true scores of the items as a network rather than a latent variable model and estimate the relations among the items (but see Epskamp et al., 2017, for possible drawbacks).

Since the posterior distributions were approximated with variational inference, the obtained posterior distributions may be biased. In general, these biases rarely affect the estimated posterior means, but the posterior variance can be underestimated (Blei et al., 2017). As a consequence, uncertainty intervals may be too narrow. To alleviate this problem, it is relatively straightforward to modify the Stan code in the appendix to use MCMC instead of variational inference (e.g., in the code in the appendix change `vb(model)` to `sampling(model)` to use MCMC). However, note that MCMC algorithms for the models discussed run for hours to obtain a reasonable number of posterior samples, whereas variational inference finishes after several minutes.

In the extended LTRM, extreme location parameters θ_{ip} are underestimated (e.g., see Figure 3.5). From a Bayesian perspective, there is little to worry about. Given the priors and the data, the posterior follows automatically. From a frequentist perspective, this bias may be worrying. This bias can be mitigated in several ways. However, we want to stress that addressing bias should be considered in light of the decisions made based on the estimates. Furthermore, bias should not be considered in isolation of the bias-variance tradeoff, that is, reducing the bias may increase the variance of an estimator, which harms generalization. For example, one straightforward approach to reduce bias is to tune the prior to minimize shrinkage. On the other hand, there are many success stories of shrinkage, Stein's paradox being a well-known example (Efron & Morris, 1977). Rather than interpreting point estimates one could instead consider the uncertainty intervals, assuming these have frequentist coverage (e.g., given enough data points or by using a procedure similar to C. Yu and Hoff (2018) or Hoff and Yu (2019)).

3.4.2 Recommendations for clinical practice

To successfully apply the extended LTRM in practice, the data should meet several minimum requirements. For instance, it should be recorded which rater gave what rating, and patient and rater covariates should contain as few missing observations as possible. Furthermore, although the model accounts for differences between raters, it is best to minimize these differences, for instance through clear scoring instructions. Minimizing differences between raters ensures that rater bias is minimal and helps to ensure validity. In addition, there should be overlap among (groups of) raters and the patients they score. That is, patients should be scored by multiple raters in such a way that there are no isolated groups of raters and patients, where one group of raters only rates one group of patients and another group of raters rates a different group of patients. A lack of overlap between two groups complicates a comparison between raters and patients between them. A lack of overlap can be avoided by having rater 1 score patients 1 through 5, having rater 2 score patients 3 through 7, etc. Additional information about patients should be added to the model, such as the reason for incarceration. That should help the extended LTRM to distinguish between groups of patients that differ on these covariates. This also holds for the raters; if certain background variables are suspected of causing rater bias then these should be included in the model.

An important step in applying any model is assessing its fit to the data. There are at least two options for doing so with the extended LTRM. First, a traditional approach is to take the residuals of the extended LTRM and examine these for any leftover structure. As in linear models, there should be no structure in the residuals if the model accurately describes the data. Second, one could compare the predictive performance of the LTRM to that of a machine learning toolbox (e.g., Random Forest or Boosting). The data set is split into a training set and a validation set. Subsequently, the models are fitted to the training set and are evaluated on the validation set. This provides an idea of how much fit is lost by using a parametric model (the extended LTRM) as opposed to a nonparametric alternative (a machine learning toolbox).

3.4.3 Conclusion

We extended the Latent Truth Rater model (LTRM) introduced by Anders and Batchelder (2015) to a model that can be applied to patients' mental health scores in forensic psychiatric hospitals. The model accounts for individual differences between raters, items, and patients. We demonstrated that the extended LTRM can provide more information about the data at hand than the raw means for two fictitious patients. In addition, we have shown that the parameters of the extended LTRM can be adequately retrieved and that the LTRM outperforms the observed mode and several machine learning toolboxes in terms of predictive power. Finally, we have provided recommendations for clinical practitioners who wish to apply the LTRM in practice. Altogether, we believe the extended LTRM constitutes a promising approach for the analysis of mental health scores in forensic psychiatric hospitals.

4

Augmenting Predictive Models in Forensic Psychiatry with Cultural Consensus Theory

Forensic psychiatric hospitals regularly monitor the mental health and forensic risk factors of their patients. As part of this monitoring, staff score patients on various items. Common practice is to aggregate these scores across staff members. However, this is suboptimal because it assumes that assessors are interchangeable and that patients are independent. An improvement over averaging scores is the use of Cultural Consensus Theory (CCT), which imposes a hierarchical model across patients, staff members, and items. While accounting for differences between patients and staff members, CCT estimates a "true" score for each patient on each item based on the consensus among staff members. Here we apply a CCT model to data from a Dutch maximum security forensic psychiatric center and use the inferences to predict violent behavior in patients. The CCT model outpredicts several alternatives, such as random forest and boosted regression trees, albeit by a small margin. We discuss practical limitations and directions for how future monitoring of patients could be adapted to maximize the added value of a CCT-based approach.

This chapter is submitted as: van den Bergh, D., Schuringa, E., & Wagenmakers, E.-J. (2023). Augmenting predictive models in forensic psychiatry with cultural consensus theory. *Manuscript submitted for publication*. 10.31234/osf.io/9kp3y

The mental health and forensic risk factors of patients in forensic psychiatric hospitals is regularly monitored with methods such as Routine Outcome Monitoring (de Beurs et al., 2011). A staff member (e.g., a clinician or psychiatrist), henceforth a *rater*, scores a patient on variety of criteria, such as problematic behavior (e.g., hostility) or protective behavior (e.g., coping skills). These scores are used to track the mental state of patients over time, to measure the effectiveness of treatment, and as a risk indicator for violent outbursts by patients.

Typically, multiple raters score each patient on different items. Standard practice is to average the scores across raters and use the averages to inform decisions. However, this is suboptimal for multiple reasons. For example, taking the average implies that raters are interchangeable and patients are independent – the raters' bias or patients' offense is not taken into account.

An improvement over averaging the scores is to use Cultural Consensus Theory (CCT; Batchelder & Anders, 2012; Batchelder & Romney, 1988; Erdfelder et al., 2020; Romney et al., 1986) to construct an appropriate model for the scores that accounts for the hierarchical structure among patients, raters, and items. In previous work, we developed such a model based on the Latent Truth Rater model (LTM; Anders & Batchelder, 2015), and demonstrated that, in theory, this model predicted better than the average and several machine learning alternatives. However, due to the lack of an empirical dataset, we could not demonstrate whether the theoretical claims hold up in practice.

Here, we apply the CCT-based model to data from a Dutch maximum-security forensic psychiatric center and use its inferences to predict whether or not a patient becomes violent. First, we briefly introduce the LTM model used and discuss two changes made compared to van den Bergh, Bogaerts, et al. (2020). Afterward, we use the LTM to augment a logistic regression. We use the augmented model to predict violent outbursts in patients and compare the predictive performance to that of frequently used machine learning models. We find that our LTM approach outperforms all other methods, albeit by a small margin. Next, we interpret the fitted model, which shows that the prior history of violence is most predictive. Finally, we discuss some practical limitations of the data set at hand and how future monitoring of patients could be adjusted to maximize the added benefit of our CCT-based approach.

4.1 Mesdag Data

The data were collected in the Dutch maximum-security Forensic Psychiatric Center Dr. S. van Mesdag between October 2016 and February 2019. The individual records were retrospectively merged into a single data set. In total, the data set contains information about 104 patients given by 188 raters on 23 items from 2 measurement occasions (18,354 observations in total). In addition to the scores on the IFTE items, the data set contains several background variables about the patients. These are age (21-30, 31-40, 41-50, or 55+ years old)[1], treatment duration (0-2 years, 2-4 years, 4-6 years, or 6+ years), diagnosis (schizophrenia and other psychotic disorders, autism spectrum disorder, Axis 1[2], personality disorder B, other personality disorders), offense (murder, arson, manslaughter, sex offense, aggravated assault, violent property crime, and moderately violent crime or a property crime), and history of violence (violent behavior 6 months before measurement 1, violent behavior 6 months before measurement 2, and violent behavior 6 months after measurement 2). When patients have multiple convictions, 'offense' indicates the most serious conviction. A distinction between personality disorder B (composed of borderline, antisocial, and narcissistic personality disorder) and other personality disorders is made because patients with personality disorder B exhibited more violent behavior.

4.1.1 IFTE

The data were collected using a Routine Outcome Monitoring instrument called the Instrument for Forensic Treatment Evaluation (IFTE; Schuringa et al., 2014, 2021). The IFTE consists of 22 items, of which 14 items are criminogenic need indicators of the Dutch risk assessment instrument HKT-R (Spreen et al., 2014), five items were designed in consultation with psychologists and psychiatrists, and three items are based on the Atascadero Skills Profile (Vess, 2001). The 22 items can be grouped into three factors: Protective behaviors, Problematic behaviors, and Resocialization Skills. The individual items are shown in Table C.1. All items are scored on a 17-point scale. Before the IFTE is scored, the rater provides their clinical judgment to answer the question: "Has the patient changed in this last period?" on a 13-point scale. This last item is treated as the 23rd item of the IFTE here. Each patient was scored with the IFTE on two separate occasions. The 17-point scale contained 5 anchor points at 1, 5, 9, 13, and 17. For example, for the item "Does the patient show problem insight?" these anchors would be 'None', 'Rarely', 'Sometimes', 'Often', and 'Always'.

[1] Ideally age is treated as a continuous variable. However, for privacy reasons, the data were anonymized, for example by categorizing age.

[2] Axis 1 is a combination of multiple disorders, such as substance-related disorders (addiction, dependence, abuse), developmental disorders (ADHD, ADD), mood disorders (depression, bipolar mood disorder), cognitive disorders (delirium, dementia, amnesia), and sexual disorders (paraphilia, pedophilia) (Segal, 2010). Note that DSM 4 categories are used as at the time of measurement, not all DSM 4 diagnoses had been converted to DSM 5 diagnoses.

4.1.2 Descriptives

Before introducing the models used and analyzing the data, we give a descriptive summary of the data. Table 4.1 shows the background characteristics of non-violent and violent patients. The raw percentages suggest that patients with violent behavior after the second IFTE measurement were also more often

Table 4.1: Characteristics of the non-violent and violent patients.

	Non-violent after T2	Violent after T2
Age		
21-30	5 (6%)	5 (21%)
31-40	40 (50%)	14 (58%)
41-50	25 (31%)	2 (8%)
51+	10 (12%)	3 (12%)
Treatment duration		
0-2 years	32 (40%)	13 (54%)
2-4 years	19 (24%)	4 (17%)
4-6 years	14 (18%)	0 (0%)
6+ years	15 (19%)	7 (29%)
Diagnosis		
Schizophrenia and other psychotic disorders	31 (39%)	9 (38%)
Autism spectrum disorder	14 (18%)	2 (8%)
Axis 1	12 (15%)	1 (4%)
Personality disorder cluster B	14 (18%)	9 (38%)
Other personality disorders	9 (11%)	3 (12%)
Offense		
Murder	10 (12%)	1 (4%)
Arson	9 (11%)	2 (8%)
Manslaughter	16 (20%)	2 (8%)
Sex offense	17 (21%)	1 (4%)
Aggravated assault	14 (18%)	10 (42%)
Violent property crime	6 (8%)	5 (21%)
Moderate violence / property crime	8 (10%)	3 (12%)
History of violence before T1		
Non-violent	67 (84%)	6 (25%)
Violent	13 (16%)	18 (75%)
History of violence before T2		
Non-violent	68 (85%)	5 (21%)
Violent	12 (15%)	19 (79%)

violent before the first measurement and in between the two measurements. For the other variables, an intuitive assessment suggests that patients with an offense of either aggravated assault, violent property crime, or moderate violence/property crime are more violent after T2 (75%) than patients with a different offense (25%).

Next, we examine the IFTE scores. Figure 4.1 shows a histogram of the raw scores across all IFTE items, raters, and patients. It is clear that the anchor points are given more often than the other scores (≈ 54% of all scores are anchors points). Furthermore, it appears that points in the middle of two anchor points are given more often than points adjacent of an anchor point (i.e., a 3 is scored more often over 2 or 4, a 7 is scored more often than 6 or 8, etc.). Given the number of patients and raters, it is evident that not all raters can have scored all patients. Figure 4.2 confirms this and shows that the rater-by-patient matrix is quite sparse.

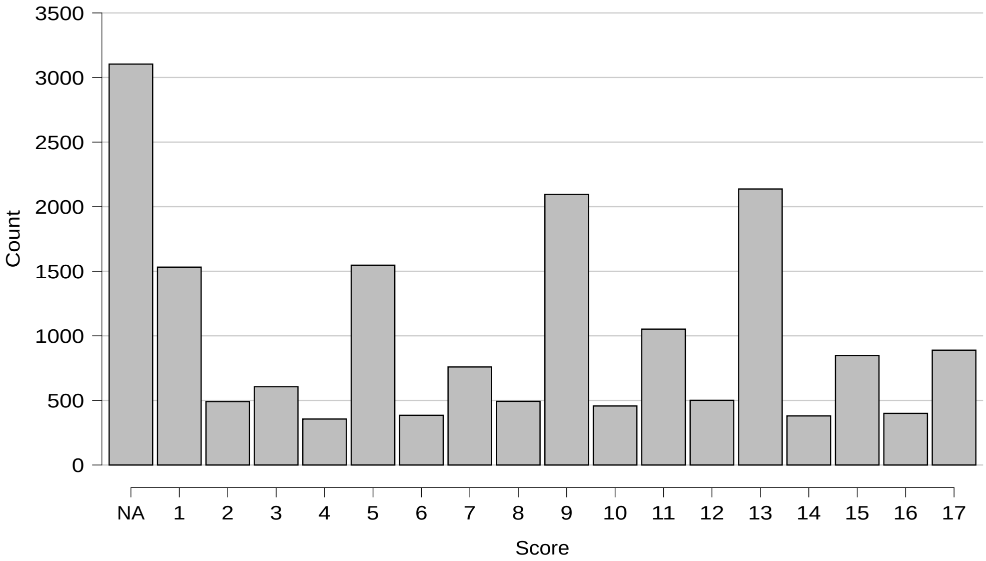


Figure 4.1: Histogram of observed scores across all patients, items, raters, and time points. The first x-axis value, NA, represents missing observations.

4.2 Cultural Consensus Theory

Cultural Consensus Theory, sometimes called "test theory without an answer key" (Batchelder & Romney, 1988), is a method to discover the "true answer" for items from the consensus among the responses. For example, suppose a patient is scored by multiple raters on hostile behavior. Multiple scores are

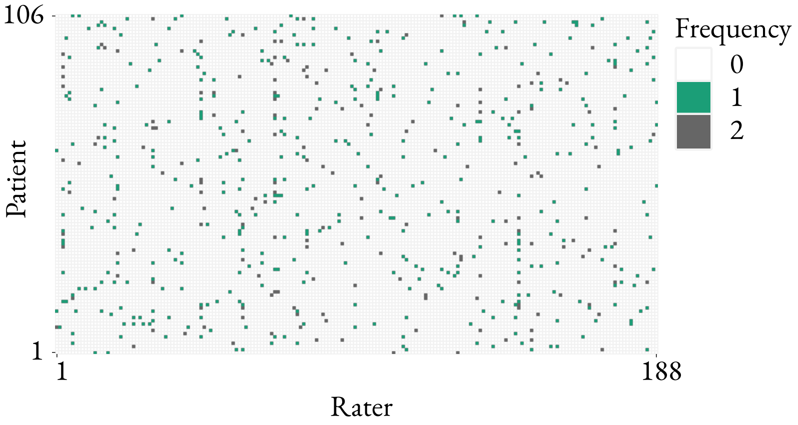


Figure 4.2: Heatmap of observed scores of rater (x-axis) against patients (y-axis).

obtained that need to be aggregated to arrive at a single score for this patient. The naive solution is to average these scores. However, as shown in Figure 4.3 averaging may lead to estimates that are severely biased.

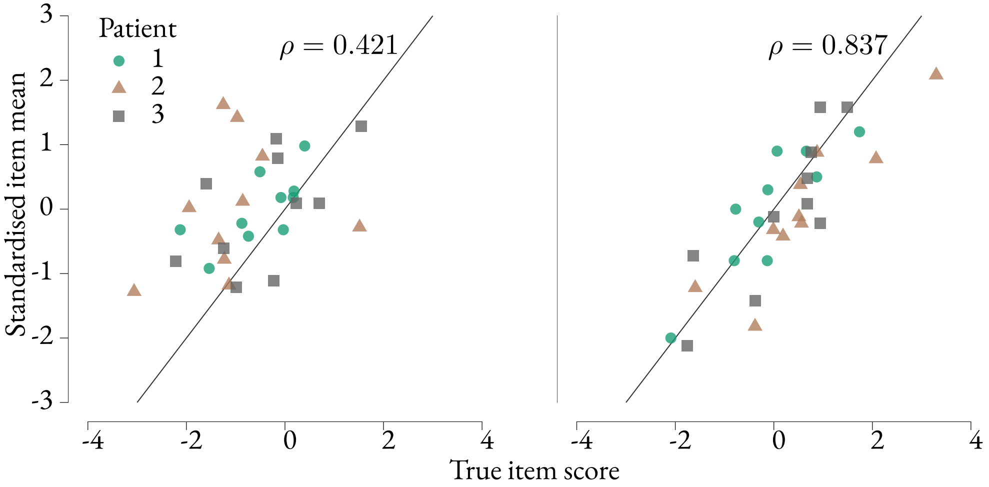


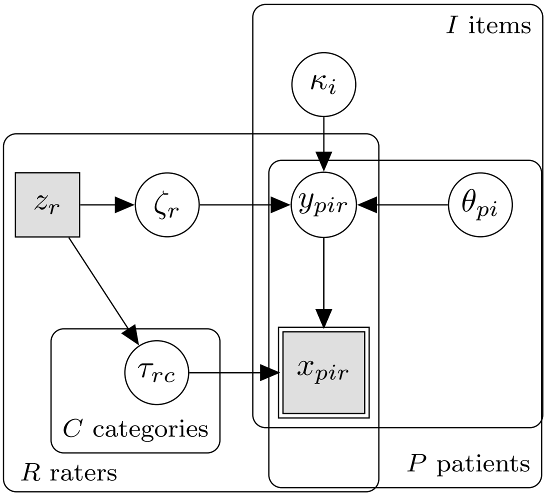
Figure 4.3: True item scores (x-axis) versus the sample mean across raters (y-axis) for three fictitious patients. The left panel shows a scenario where the raters are heterogeneous and consequently the performance of the sample mean is poor. In the right panel, the raters are homogeneous and the sample mean performs much better.

The average score disregards all additional information that is available. It ignores the individual differences between raters, for example, this assumes that all psychiatrists score hostility in the same way, and it ignores group differences among raters; for example, there is no difference in scores by psychiatrists as opposed to clinicians, or other staff members. In addition, the average ignores any additional information about the patient at hand, such as the nature of the offense and the diagnosis.

Cultural consensus theory provides a model-based framework for pooling information from multiple raters to form a consensus (Anders et al., 2014). There exist a variety of CCT models, each applicable to different types of data. For example, the General Condorcet model (Batchelder & Romney, 1986) applies to dichotomous data, the Continuous Response model (Anders et al., 2014) is suited for continuous data, and the Latent Truth Rater model (Anders & Batchelder, 2015) is suited for ordinal data. As the IFTE scores are ordinal, we use the Latent Truth Rater model to analyze the Mesdag data.

4.2.1 The Latent Truth Rater Model

The Latent Truth Rater Model (LTM) is a CCT model for ordinal data (Anders & Batchelder, 2015). Previously, we extended the LTM to handle data from multiple patients (van den Bergh, Bogaerts, et al., 2020) and Figure 4.4 shows the LTM for multiple patients.



$$x_{pir} = \begin{cases} 1 & \text{if } y_{pir} \le \delta_{r1} \\ c & \text{if } \delta_{r,c-1} < y_{pir} \le \delta_{rc} \\ C & \text{if } y_{pir} > \delta_{r,C-1} \end{cases}$$

$$\begin{aligned} y_{pir} &\sim \text{Logistic}\left(\theta_{pi},\ \kappa_i/\zeta_r\right) \\ \delta_{rc} &\sim \text{Normal}\left(0,\ 1\right) \\ \theta_{pi} &\sim \text{Normal}\left(\mu_\theta,\ \sigma_\theta^2\right) \\ \kappa_i &\sim \text{Gamma}\left(\mu_\kappa^2/\sigma_\kappa^2,\ \mu_\kappa/\sigma_\kappa^2\right) \\ \zeta_r &\sim \text{Gamma}\left(\mu_{\zeta_r}^2/\sigma_{\zeta_r}^2,\ \mu_{\zeta_r}/\sigma_{\zeta_r}^2\right) \end{aligned}$$

Figure 4.4: Graphical model of the Latent Truth Rater Model for multiple patients. Note that the thresholds $\boldsymbol{\delta}_r$ are constrained to be ordered, that is, for all raters we have $\delta_{r1} \le \cdots \le \delta_{rc} \le \cdots \le \delta_{r,C-1}$.

Here, x_{pir} is the observed score given to patient p on item i by rater r. This score is assumed to be deterministically generated from a continuous latent appraisal y_{pir} that is discretized to an ordinal scale by the thresholds δ_{rc}. In particular, we have that

$$x_{pir} = \begin{cases} 1 & \text{if } y_{pir} \leq \delta_{r1} \\ c & \text{if } \delta_{r,c-1} < y_{pir} \leq \delta_{rc} \\ C & \text{if } y_{pir} > \delta_{r,C-1} \end{cases}$$

Since the appraisal score is latent, the deterministic function above implies the following probabilistic model over the observed scores:

$$P(x_{pir} \mid y_{pir}, \boldsymbol{\delta}_r) = \begin{cases} 1 - F\left(y_{pir} - \delta_{r1}\right) & \text{if } x_{pir} = 1, \\ F\left(y_{pir} - \delta_{r,c-1}\right) - F\left(y_{pir} - \delta_{rc}\right) & \text{if } 1 < x_{pir} < C, \\ F\left(y_{pir} - \delta_{r,C-1}\right) & \text{if } x_{pir} = C. \end{cases}$$

where $F\left(\right)$ is the logistic cumulative distribution function.[3]

Next, we explain how the latent appraisals and thresholds come about. The appraisals are drawn from a logistic distribution with location θ_{pi}, the true score for patient p on item i. The scale of the logistic distribution is the ratio of the item difficulty κ_i to the rater competence ζ_r. A higher item difficulty means that the appraisals are more noisy, which leads to a more dispersed probability distribution over possible scores. Conversely, a higher rater competence means that the appraisals are less noisy, which leads to a more concentrated distribution over the outcomes. There are $C - 1$ ordered thresholds for each rater, which are assigned a standard normal prior for identification purposes.

There are two differences in the model specification above compared to our previous work (van den Bergh, Bogaerts, et al., 2020). First, we previously modeled the thresholds using two rater-specific parameters. However, in simulations, we noticed that these two parameters provide too little flexibility when the ordinal scale consists of 17 categories and has a multimodal distribution (see Figure 4.1), as in the Mesdag data. Therefore we decided to model the thresholds individually. This modeling choice complicates the interpretation of the differences between the thresholds across raters; however, that is also not the goal of this paper. Second, we previously allowed the item difficulty parameter to vary across patients, which allows for the possibility that some items may be more difficult or easy to assess for particular patients (e.g., some patients may cooperate more than others). To estimate this patient-item interaction there must be a sufficient number of raters who score each patient. However, when simulating data with a rater-to-patient ratio similar to that in the data at hand, we noticed that there are simply too few observations to

[3] The choice of the distribution function is arbitrary, in principle it is possible to use any continuous cumulative distribution function.

reliably estimate the deviations in item difficulty across patients. Therefore, we only vary item difficulty across items and not across patients.

Altogether, the LTM extracts a true score (θ_{pi}) for each patient on every item while accounting for the hierarchical structure among patients, raters, and items. The posterior distributions for the true scores are the basis for all subsequent analyses.

4.2.2 Augmenting Logistic Regression with the LTM

In the next step, we use logistic regression to predict violent behavior, where we use the estimated true scores (θ_{pi}) from the LTM as additional predictors. We do so in a fully Bayesian approach, that is, we constructed a joint model for the violent behavior and the patient ratings.[4] Figure 4.5 shows a graphical model of the logistic regression combined with the LTM. The latent truth for each patient on each item is seen as a covariate in the logistic regression model. Here, v_p denotes the violent behavior of patient p. The regression coefficients of the true scores are denoted η_i. In addition to the true scores, we regress the background variables z_{pd} with associates coefficients η_d onto violent behavior.

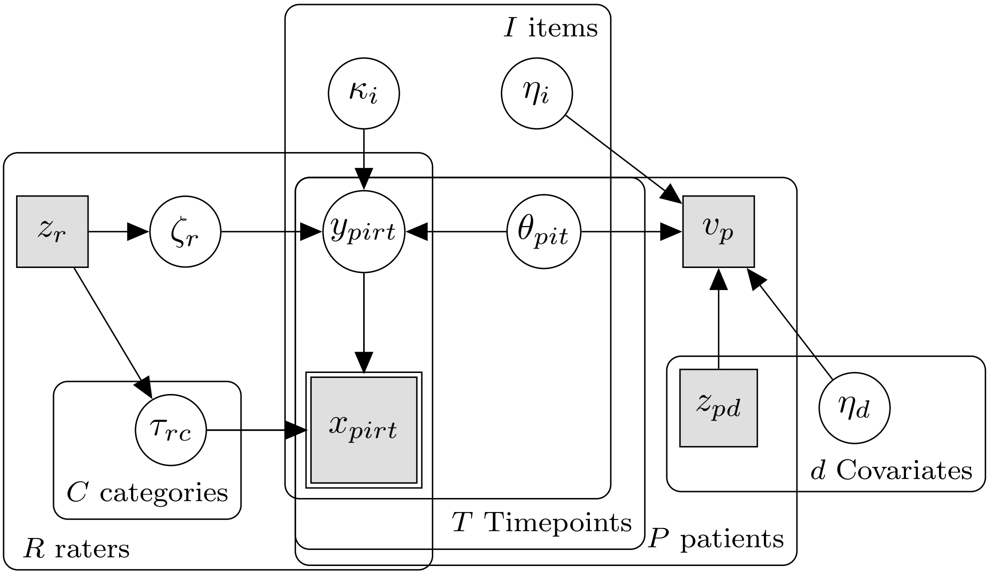


Figure 4.5: Graphical model of Logistic Regression Augmented with the Latent Truth Rater Model. The true scores

[4] An alternative is a two-step approach where in the first step the LTM is fit to the patient ratings. In the second step, the estimates of the LTM (e.g., the posterior means) are used as predictors in a logistic regression model. While this is computationally faster, it also ignores the uncertainty in the analysis of the patient ratings.

The index t indicates the measurement occasion (1 or 2). Rather than fitting the LTM completely anew for each time point, we assume that all rater parameters (κ_r and $\boldsymbol{\delta}_r$) and item parameters (κ_i) constant but allowed the patient parameters to differ.

4.2.3 Implementation

We estimated the parameters of the LTM and the combined LTM-Logistic regression model using a Bayesian approach. To explore the posterior distributions of the model parameters we used Stan (Carpenter et al., 2017). Rather than Markov chain Monte Carlo (MCMC) we used variational inference, which was computationally faster while providing similar results in terms of parameter retrieval and model predictions (Kucukelbir et al., 2017). All analyses were done using R (R Core Team, 2022) and Stan models were run using the R package `cmdstanr` (Gabry & Češnovar, 2022). The code for the analyses is available in the online appendix at https://github.com/vandenman/CCT-Logistic. Although the data cannot be shared due to privacy concerns, the repository contains simulated data that can be used to run the code.

The machine learning methods and the logistic regression models that will be introduced in the next section cannot handle missing values in the predictors. Therefore, we imputed the missing values using the R package mice (van Buuren & Groothuis-Oudshoorn, 2011). In the Bayesian analyses, we marginalized out the missing values.

4.3 Predictive Performance

Here we compare the predictive performance of the logistic regression model that is augmented with the LTM (LR-LTM) to several reasonable alternatives and three baseline models. The first baseline model is an intercept-only logistic regression (LR-Intercept). Comparing the predictive performance of a model with the LR-Intercept constitutes a sanity check that a model outperforms the baseline prevalence of violence. The second baseline model we use is a logistic regression model with all covariates but not the IFTE items (LR-No IFTE). As a third baseline model, we use a logistic regression model with only prior violence as predictors (LR-Violence). As more plausible competing models, we consider logistic regression (LR-All), random forest, and boosted regression trees (GBM) which are trained with history of violence, patient covariates, and IFTE scores. These three competing models have in common that they are designed for purely rectangular data. That is, each row of the data set contains one outcome (violent or nonviolent behavior) and a number of predictors. However, the raw data of the IFTE contains repeated observations, since patients were rated multiple times by different raters. To accommodate this we (naively) average across different raters to obtain a single score for each item and time point.

Altogether, by comparing these seven models we aim to answer the following three questions: (1) Can predictive models outperform the base prevalence of violent behavior? (2) Do models with the IFTE scores perform better than the baseline models without the IFTE scores? (3) Does the LR-LTM perform better than the models that naively average across raters?

To examine predictive performance we used 10-fold stratified cross-validation. Each fold consisted of 8 nonviolent observations and 2 or 3 violent observations. We quantified model performance with prediction accuracy and the Brier score (Brier et al., 1950). Prediction accuracy is defined as the fraction of correct predictions. We converted model probabilities for violent or non-violent behavior into binary predictions by comparing them with 0.5. In other words, if a model prediction for observation i, $\hat{y}_i \in [0, 1]$, is larger than 0.5 then the predicted label is violent, otherwise, it is non-violent. The Brier score is defined as $N^{-1} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$, i.e., the mean squared error between the observed labels, $y \in \{0, 1\}$ and the model predictions.

Table 4.2 shows the prediction accuracy and the Brier score averaged over the 10-folds. The LR-LTM has the best classification of violence and the lowest Brier score, and the LR-Violence performs second best. The difference in classification performance is $0.884 - 0.866 = 0.018$, which for a data set of 104 patients implies that the LR-LTM makes more accurate predictions than the LR-Violence for about 2 patients. Possibly striking is the poor performance of the LR-All. The standard logistic regression model clearly suffers from overfitting, as indicated by the high training performance but poor test performance. This result makes sense as the standard logistic regression model does not do anything special to combat overfitting, unlike the machine learning alternatives or Bayesian logistic regression used by the LR-LTM.

Figure 4.6 show the Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) averaged across cross-validation runs for all methods except the LR-Intercept.[5] For each cross-validation run, we computed the true positive rate and false-positive rate with the same set of thresholds run and afterward we averaged these. The AUC was obtained by averaging the AUCs of each individual cross-validation, rather than computing the AUC for the averaged ROC curve. In line with the previous results, the LR-LTM performs best and the LR-Violence method performs second best.

To summarize, it is evident that all models, except for logistic regression, outperform the intercept-only baseline model. Furthermore, the results show that the LTM augmented logistic regression model performs best, albeit by a

[5] The ROC for the intercept-only model is by definition the identity function with an area under the curve of 0.5.

Table 4.2: Predictive performance of violent behavior. The values are the average of 10 cross-validations; the standard deviation is shown in parentheses.

	Classification		Brier score	
Method	Train	Test	Train	Test
LR-LTM	0.980 (0.008)	0.884 (0.078)	0.047 (0.006)	0.095 (0.028)
LR-Violence	0.865 (0.007)	0.866 (0.063)	0.100 (0.003)	0.107 (0.030)
LR-No IFTE	0.948 (0.034)	0.837 (0.105)	0.032 (0.020)	0.142 (0.109)
Random forest	0.996 (0.006)	0.835 (0.081)	0.033 (0.002)	0.114 (0.040)
GBM	0.880 (0.011)	0.836 (0.066)	0.093 (0.003)	0.125 (0.023)
LR-All	1.000 (0.000)	0.727 (0.182)	0.000 (0.000)	0.258 (0.173)
LR-Intercept	0.769 (0.004)	0.771 (0.038)	0.177 (0.002)	0.177 (0.020)

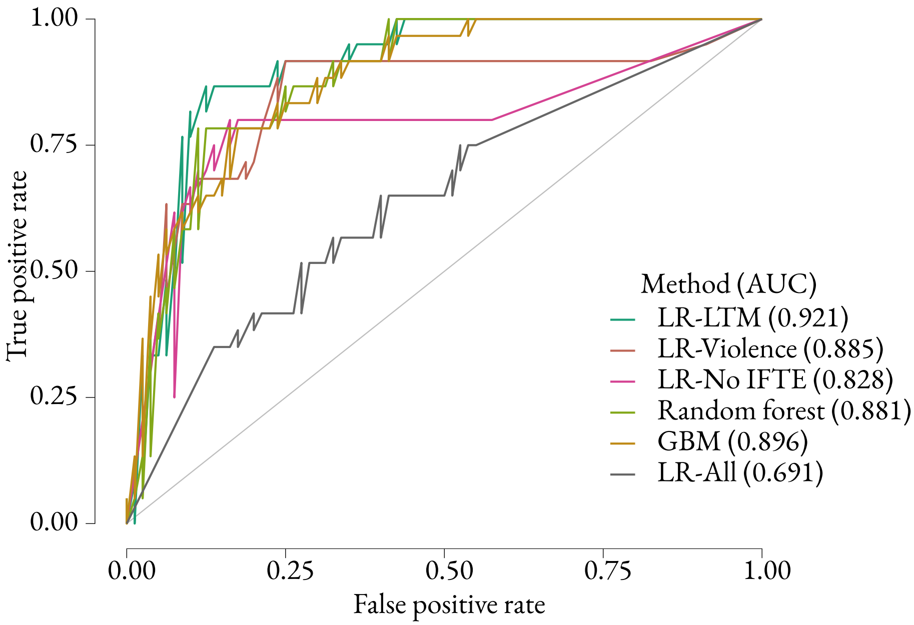


Figure 4.6: ROC curves for the considered methods. The legend shows the area under the curve in parentheses.

small margin. The logistic regression model with only violence (LR-Violence) outperformed the two machine learning alternatives that naively averaged the scores from the IFTE (GBM and Random forest). This indicates that the IFTE has added value for prediction, but only if it is analyzed properly.

4.4 Interpretation of the LTM

Now that we have shown that the LTM had adequate predictive performance, we interpret the model parameters. Figure 4.7 shows the posterior means and 95% credible intervals for the LTM fitted to the complete data. Prior history of violence has the largest coefficients, whereas the other coefficients all appear close to zero. As Table 4.1 already indicated, the history of violence has a large

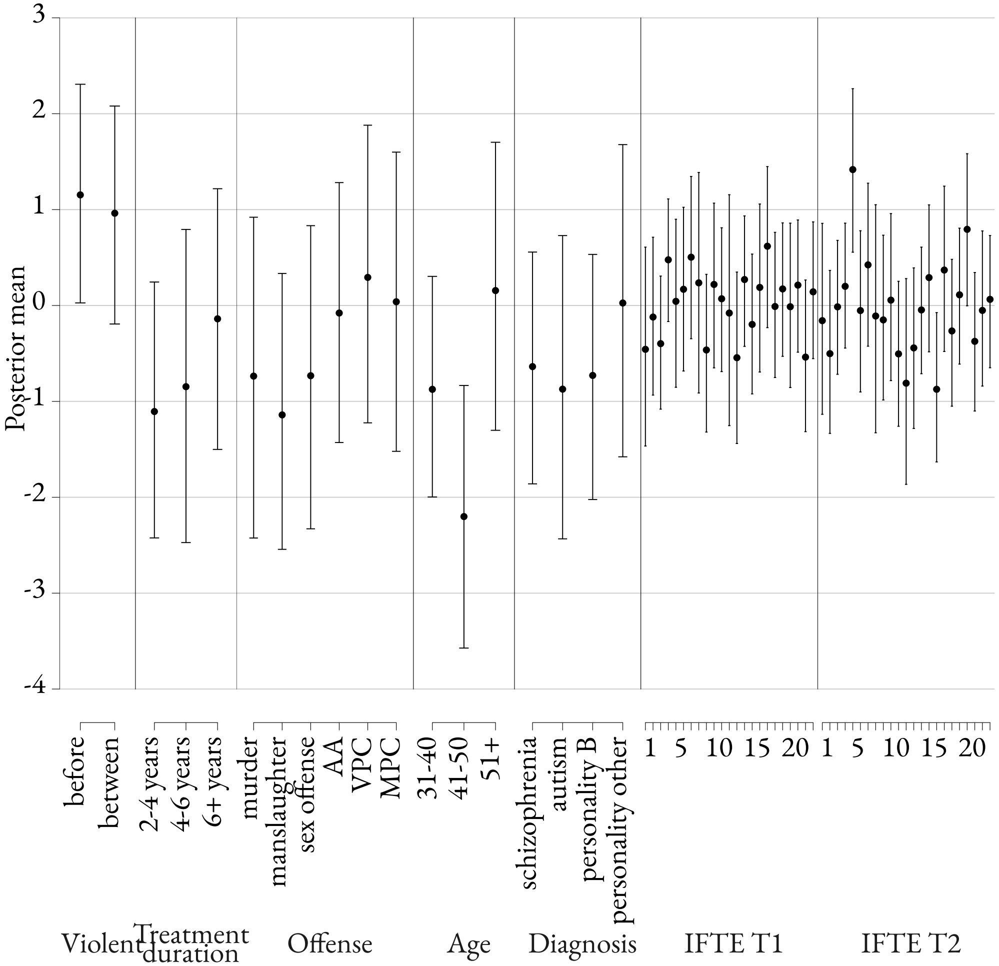


Figure 4.7: Posterior means and 95% credible intervals for the coefficients of the logistic regression model (η_d and η_i in Figure 4.5). The reference categories are treatment duration 0-2 years, diagnosis Axis 1, and offense arson. The abbreviations VPC, MPC, and AA stand for violent property crime, moderately violent property crime, and aggravated assault.

effect. In contrast, the influence of the IFTE items is less clear as all coefficients appear to be close to zero. However, while the other coefficients shown relate the probability of violent behavior to observed data (i.e., a fixed value), the coefficients for the IFTE relate the probability of violent behavior to the latent truth θ_{pi} (i.e., a probability distribution). Therefore it is difficult to interpret the contribution of the IFTE without also considering the latent truth of the patients. Figure 4.8 visualizes the impact of the IFTE as a whole by visualizing the posterior mean of the logit of the probability of violent behavior summed over all IFTE items for each patient. For a particular patient p this logit is obtained by computing $\sum_{t=1}^{2}\sum_{i=1}^{23}\theta_{pit}\eta_{it}$ averaged across all posterior samples. For most patients that showed violent behavior after T2, the posterior mean

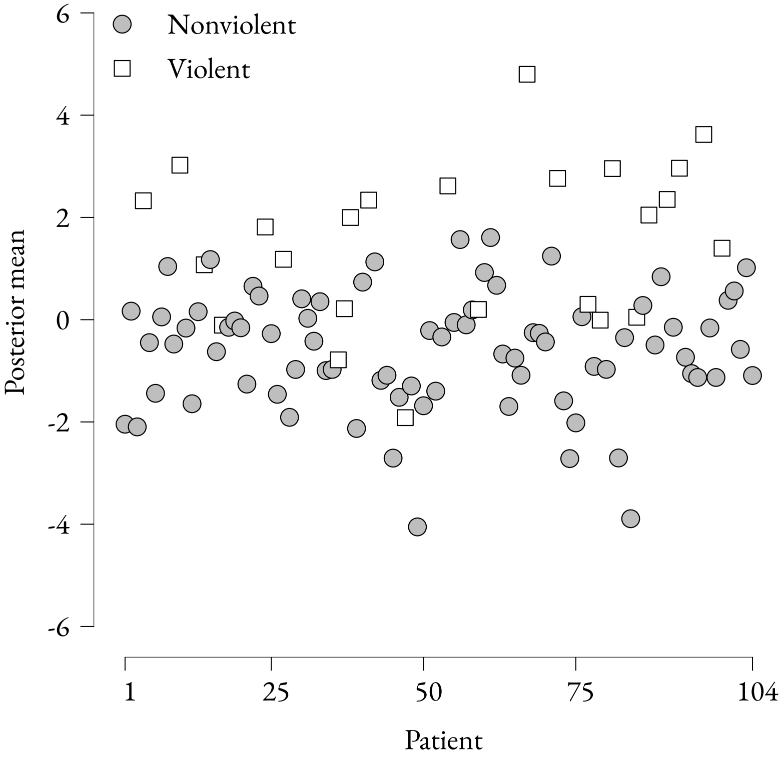


Figure 4.8: Posterior mean of the effect of the total impact of the IFTE items for each patient $\left(\sum_{t=1}^{2}\sum_{i=1}^{23}\theta_{pit}\eta_{it}\right)$. Shape and color indicate whether patients showed violent behavior after T2 (white squares) or not (gray circles).

of the impact of the aggregated IFTE is larger than 1, which is about the posterior mean for the effect of prior violence. However, the uncertainty of these estimates is large, and the average 95% credible interval ranges from -8 to 8.[6] Nevertheless, Figure 4.8 shows that the IFTE matters for prediction, but also that its added value differs across patients.

With the model comparisons in the previous section and parameter estimates in the previous paragraph we have shown that IFTE scores analyzed with the LTM provide added value to predict aggressive behavior. In contrast, the sample means of the IFTE items did not seem to improve predictions of violent behavior. This indicates that for this data set the sample mean is not a good approximation to the latent truth as estimated by the LTM. As shown with simulated data in Figure 4.3, the relationship between the latent truths and item sample means depends strongly on the heterogeneity of the raters. Figure 4.9 visualizes the posterior mean of the latent truth θ_{pi} against the standardized sample means for each patient and item. The latent truths and

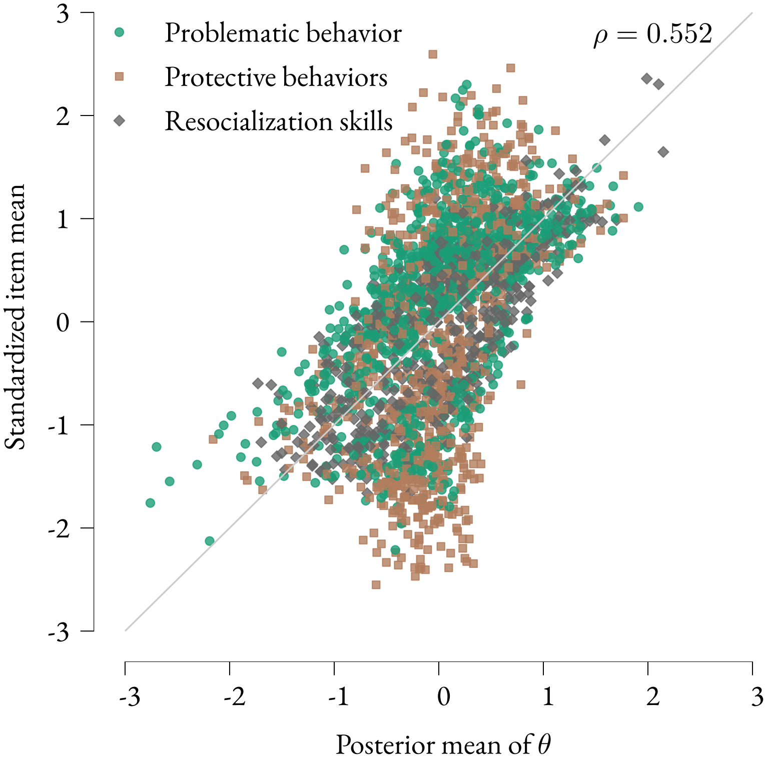


Figure 4.9: Posterior mean of the true item scores (x-axis) versus the sample mean across raters (y-axis). Colors indicate the item factors as listed in Table C.1. The correlation is around 0.5 which suggests that there is substantial variability among raters that is ignored by the sample mean.

[6] The large uncertainty here is inevitable, as the variance of a sum of random variables is often close to the sum of the variances of the individual parameters.

the sample means are somewhat correlated, but it is clear that the sample means are not a good approximation to the latent truths which suggests that there is substantial variability among raters that is accounted for by the LTM but ignored by the sample mean.

4.5 Discussion

Here we applied a Cultural Consensus Theory model to scores of patients in a Forensic Psychiatric Center. We used this CCT model to augment the predictive performance of a logistic regression model and showed that our CCT-infused logistic regression model outpredicted all other candidate models. Interpreting the influence of individual items of the IFTE is not straightforward. However, the aggregated effect of the IFTE appeared substantial but also varied across patients. Nevertheless, the uncertainty in the parameter estimates is too large to warrant strong conclusions and interpretations, other than the unsurprising result that prior history of violent behavior is a strong predictor of future violence.

Here we devised a joint model for the IFTE scores and prediction of violent behavior. An alternative approach is to first fit a CCT model to the IFTE scores and then in a second step use its estimates for prediction. A disadvantage of this is that it ignores the uncertainty in the CCT estimates. On the other hand, a pragmatic advantage is that it decreases the running time of the models. Furthermore, a two-step approach also opens up the possibility to use machine learning methods for prediction. While such an approach has shown merit before (see e.g., Plonsky et al., 2017), it is probable that using a machine learning model would further complicate the interpretation of the model parameters and predictions which may have the unintended effect that in practice the recommendations are not adopted because the model cannot be fully understood.

4.5.1 Suggestions for Future Data Collection

The interpretation of the parameters of the LTM in this application is hindered by the use of unconstrained thresholds. However, this modeling decision was mandated by the structure of the data, which, due to the large number of response categories, showed patterns that could not be described by a simpler function for the thresholds (e.g., a preference for anchor points). These response patterns were somewhat unexpected, as earlier studies did not observe such patterns (e.g., see Figure 2 of Schuringa et al., 2014). For future data collection, it would be worthwhile to explore options to make the distribution of response scores unimodal. This could be done through more clear instructions for the raters, or by collapsing infrequently used response categories. Such measures would facilitate the modeling of the data and simultaneously reduce the likelihood of spurious patterns in the data.

The frequency of measurements is key to the usefulness of the IFTE scores. Schuringa et al. (2019) previously found that the most recent measurement is most predictive, which indicates that the scores' predictive value likely decreases over time. Rather than scoring patients every 6 months, as in the data at hand, it would make more sense to rate them every few weeks. Although it may be practically difficult to score patients regularly, there are opportunities to use self-reports for this purpose (Bousardt et al., 2016; Tuente et al., 2021). A downside of self-reports is that they may introduce additional variance due to the lack of standardized scoring, or meaningless responses in case of non-cooperation.

4.5.2 Limitations

A limitation of our approach is that patients' violent behavior was collapsed into three discrete time points and consequently also modeled as such. In reality, however, not all patients show violent behavior at the same time, but rather there is variability in when the violent behavior occurs. To properly describe this a continuous time series approach would be required, for example, by adding autoregressive components. One benefit of doing so is that it becomes possible to predict the risk of violent behavior at different points in the future. For example, a continuous time series approach could predict the risk of violent behavior occurring next week, as opposed to a time-independent prediction that our current model makes. Another benefit is that it becomes possible to use time-varying covariates. For example, patients may show less or more violent behavior during holidays or on their birthdays.

Furthermore, we did not account for the structure of the IFTE items. Each item of the IFTE is designed to load on a particular factor, see Table C.1. This structure could be added to the model introduced here as a confirmatory factor model. This would imply that the true scores θ_{pi} load on their common factors, and that the factor scores are used to predict violent behavior. If the factor model fits well then it could facilitate the model interpretation, as this can then be done on the level of the factor rather than the items. However, it is unlikely that this would improve predictive performance. On the other hand, if the factor model does not fit well then predictive performance would suffer.

4.5.3 Conclusion

We applied the LTM introduced by Anders and Batchelder (2015) and adapted previously in van den Bergh, Bogaerts, et al. (2020) to data of patients in a

Forensic Psychiatric Center. We showed that including the IFTE items slightly improves predictive performance, but only if the scores from different raters are analyzed properly and not when the scores of different raters are averaged. We discussed different approaches to extend the models and to make the data more informative.

Part II

Bayesian Model Averaging

4

5

A Cautionary Note on Estimating Effect Size

An increasingly popular approach to statistical inference is to focus on the estimation of effect size. Yet, this approach is implicitly based on the assumption that there is an effect while ignoring the null hypothesis that the effect is absent. We demonstrate how this common "null hypothesis neglect" may result in effect size estimates that are overly optimistic. As an alternative to the current approach, a "spike-and-slab" model explicitly incorporates the plausibility of the null hypothesis into the estimation process. We illustrate the implications of this approach and provide an empirical example.

This chapter is published as: van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021b). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245921992035.

Consider the following hypothetical scenario: a colleague from the biology department has just conducted an experiment and approaches you for statistical advice. The analysis yields $p < 0.05$ and your colleague believes that this is grounds to reject the null hypothesis. In line with recommendations both old (e.g., Grant, 1962; Loftus, 1996) and new (e.g., Cumming, 2014; Harrington et al., 2019) you convince your colleague that it is better to replace the p-value with a point estimate of effect size and a 95% confidence interval (but see Morey, Hoekstra, et al., 2016). You also manage to convince your colleague to plot the data (see Figure 5.1). Mindful of the reporting guidelines of the *Psychonomic Society*[1] and *Psychological Science*[2], your colleague reports the result as follows: "Cohen's $d = 0.30$, CI $= [0.02, 0.58]$".

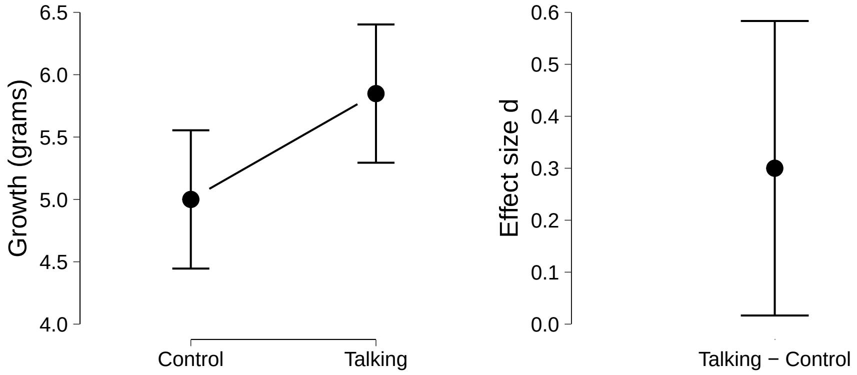


Figure 5.1: Standard estimation results for the fictitious plant growth example. Left panel: a descriptives plot with the mean and 95% confidence interval of plant growth in the two conditions. Right panel: point estimate and 95% confidence interval for Cohen's d.

Based on these results, what would be a reasonable point estimate of effect size? A straightforward and intuitive answer is "0.30". However, your colleague now informs you of the hypothesis that the experiment was designed to assess: "plants grow faster when you talk to them".[3] Suddenly, a population effect size of "0" appears eminently plausible. Any observed difference may merely be due to the inevitable sampling variability.

The example above is rhetorical but serves to underscore the potential conflict between standard reporting guidelines and common sense. The example raises the question: When are effect sizes overestimated? Standard point esti-

[1] https://www.springer.com/psychology?SGWID=0-10126-6-1390050-0

[2] https://www.psychologicalscience.org/publications/psychological_science/ps-submissions#STAT

[3] This example is inspired by Berger and Delampady (1987).

mates and confidence intervals ignore the possibility that the effect is spurious (i.e., the null hypothesis $\mathcal{H}_0$). This is not problematic when $\mathcal{H}_0$ is deeply implausible, either because $\mathcal{H}_0$ was highly unlikely *a priori* or because the data decisively undercut $\mathcal{H}_0$. But when the data fail to undercut $\mathcal{H}_0$, or when $\mathcal{H}_0$ is highly likely *a priori* (i.e., "plants do not grow faster when you talk to them"), then $\mathcal{H}_0$ is not ruled out as a plausible account of the data. Effect size estimates that ignore a plausible $\mathcal{H}_0$ are generally overly optimistic and overly confident: the fact that $\mathcal{H}_0$ provides an acceptable account of the data should shrink effect size estimates towards zero. The statistical benefits of shrinkage are described in Efron and Morris (1977; see also Davis-Stober et al., 2018; Rouder and Lu, 2005; Shiffrin et al., 2008); the benefits of shrinking estimates towards zero are discussed for instance in George and McCulloch (1993) and Iverson et al. (2010), and van Erp et al. (2019).

The above point estimate, "0.30", may seem purely data-driven, but it is based on a model that assumes an effect size different from zero. In this paper we propose an alternative model to estimate effect size: the so-called "spike-and-slab" model. First, we formally introduce the spike-and-slab model. Second, we apply the spike-and-slab model to the example in the introduction and illustrate how it tempers the estimated effect size. Third, we visualize how the spike-and-slab model may shrink the estimated effect size toward zero in general. Fourth, we demonstrate the spike-and-slab model by reanalyzing the data of Heycke et al. (2018). Finally, we conclude with practical recommendations and a discussion on when to use the spike-and-slab model.

5.1 A Spike-and-Slab Perspective

The spike-and-slab approach has been widely discussed in the statistical literature (e.g., Clyde et al., 1996; Geweke, 1996; Ishwaran, Rao, et al., 2005; Mitchell & Beauchamp, 1988; O'Hara, Sillanpää, et al., 2009) and in the psychological literature (e.g., Bainter et al., 2020; Iverson et al., 2010; Rouder et al., 2018; C.-H. Yu et al., 2018). Conceptually, the approach is relatively straightforward.

As usual, the statistical goal is to infer the population effect size from a set of sample observations. Let δ denote the population effect size, let $\hat{\delta}$ denote a point estimate, and let $\hat{\delta} \mid \mathcal{H}_1$ denote a point estimate assuming the alternative hypothesis, $\mathcal{H}_1$. Assuming the null hypothesis $\mathcal{H}_0$ leads to $\hat{\delta} \mid \mathcal{H}_0$, which usually equals 0. Key is that both estimates, $\hat{\delta} \mid \mathcal{H}_1$ and $\hat{\delta} \mid \mathcal{H}_0$, are *conditional* on the hypotheses. For example, $\hat{\delta} \mid \mathcal{H}_1$ should be read as "the estimated effect size under the alternative hypothesis that the effect exists". To the best of our knowledge, all existing guidelines for reporting effect size estimates recommend that researchers provide $\hat{\delta} \mid \mathcal{H}_1$; implicitly, the guidelines suggest to ignore $\mathcal{H}_0$, resulting in the notion that the population effect size is nonzero. In contrast, in the spike-and-slab model, the estimate of effect size is determined by both $\mathcal{H}_1$ and $\mathcal{H}_0$.

As the name suggests, the spike-and-slab model consists of two components. The first component, the spike, corresponds to the position that talking to plants does not affect their growth (i.e., $\delta = 0$), whereas the second component, the slab, corresponds to the position that speaking to plants does affect their growth (i.e., $\delta \neq 0$). The spike and slab are analogous to $\mathcal{H}_0$ and $\mathcal{H}_1$ discussed above. Both components are commonly deemed *a priori* equally likely, such that the prior probability for each component is $1/2$.

One can assign prior probabilities other than $1/2$, if this is motivated by prior research, prior data, or existing theories (e.g., B. M. Wilson & Wixted, 2018). After observing the data, the prior probabilities of both components, $\Pr(\text{spike})$ and $\Pr(\text{slab})$, are updated to posterior probabilities, $\Pr(\text{spike} \mid \text{data})$ and $\Pr(\text{slab} \mid \text{data})$.

By applying the spike-and-slab model we learn about the relative plausibility of the two components; in addition, the spike-and-slab model produces a *marginal* estimate of effect size – a weighted combination of effect sizes from the spike and from the slab (for mathematical detail see the online Appendix). In other words, the spike-and-slab model yields an overall effect size averaged across the spike and the slab, with averaging weights determined by the respective posterior probabilities:

$$\hat{\delta} = \left(\hat{\delta} \mid \text{spike}\right) \Pr(\text{spike} \mid \text{data}) + \left(\hat{\delta} \mid \text{slab}\right) \Pr(\text{slab} \mid \text{data}) . \tag{5.1}$$

Marginalizing across model components according to their posterior plausibility is a uniquely Bayesian operation, and this is the statistical framework we adopt in this paper (for an accessible introduction to Bayesian inference see Vandekerckhove et al., 2018). Researchers who prefer a frequentist approach can accomplish shrinkage by using penalized maximum likelihood methods such as LASSO and ridge regression (Tibshirani et al., 2005). Another option open to frequentists is to marginalize across the spike and the slab for instance by using the Akaike Information Criterion (AIC; Akaike, 1973) and defining the averaging weights as follows. Let $\Delta\text{AIC} = (\text{AIC} \mid \text{spike}) - (\text{AIC} \mid \text{slab})$, the difference in AIC between the spike and the slab. Next we use the "Akaike weight" w_{spike} as a substitute for the posterior probability of the spike: $w_{\text{spike}} = \exp(-1/2\,\Delta\text{AIC}) / (1+\exp(-1/2\,\Delta\text{AIC}))$ (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). The substitute for the posterior probability of the slab is simply: $w_{\text{slab}} = 1 - w_{\text{spike}}$.

Note that when the spike is located at $\delta = 0$, as is usually the case, then $\left(\hat{\delta} \mid \text{spike}\right) \Pr(\text{spike} \mid \text{data}) = 0$, and consequently Equation 5.1 simplifies to

$$\hat{\delta} = \left(\hat{\delta} \mid \text{slab}\right) \Pr(\text{slab} \mid \text{data}) . \tag{5.2}$$

This equation shows that the spike-and-slab estimate $\hat{\hat{\delta}}$ equals the estimate that is generally recommended in reporting guidelines, $\left(\hat{\delta} \mid \text{slab}\right)$, but reduced by the posterior probability for $\mathcal{H}_1$. This shrinkage towards zero becomes negligible when the posterior probability for $\mathcal{H}_1$ approaches 1.

To illustrate both the overestimation and the spike-and-slab model we reanalyze the fictitious data from Figure 5.1. R code for the analysis is available at https://osf.io/uq8st/. Remember that the frequentist point estimate for the effect size conditional on $\mathcal{H}_1$, or the slab, was $\hat{\delta} = 0.30$, with a confidence interval of 95% CI: [0.02, 0.58]. The Bayesian equivalent is $\hat{\delta} = 0.29$, with a credible interval of 95% CRI: [0.02, 0.57]. Figure 5.2 contrasts this Bayesian slab-only estimate against the spike-and-slab estimate.

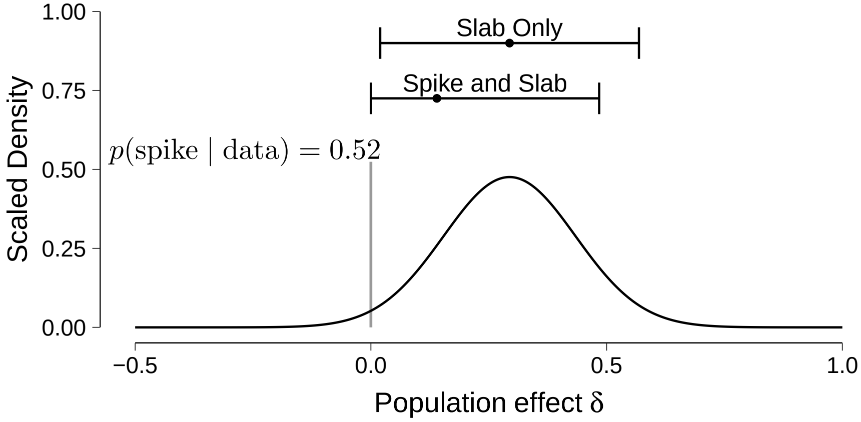


Figure 5.2: The spike-and-slab model. The black line represents the posterior distribution of effect size given the slab (i.e., the effect is non-zero). The posterior is scaled so that its mode ($\hat{\delta} = 0.29$) equals the posterior probability of the alternative model (i.e., $p(\text{slab} \mid \text{data}) = 0.48$). The grey line represents the posterior probability of the spike (i.e., $\hat{\delta} = 0$: the effect is absent). The error bars and dots above the density show 95% credible intervals and the posterior mean for the slab-only model and for the spike-and-slab model.

Compared to the traditional results based only on the slab, the posterior mean and central 95% credible interval of the spike-and-slab model are shrunken towards 0 (i.e., 0.14, 95% CRI: [0.00, 0.48] vs. 0.29, 95% CRI: [0.02, 0.57]). This shrinkage is due to the non-negligible probability that the effect is absent. Here, the posterior probability of the spike after seeing the data, 0.52, is almost identical to its prior probability. In the figure, the plausibility that the effect is absent is represented by the height of the spike, and the uncertainty about the effect's magnitude, given that it is present, by the width of the slab. Note that if the posterior probability of the spike was reduced, the spike-and-slab results would approach those of the slab-only model.

5.2 The Influence of the Spike

In the fictitious example, the spike-and-slab model reduces the estimated effect size by shrinking estimates of effect size towards zero. The result may not be surprising, as the effect was small. However, it makes one wonder to what extent the spike-and-slab model helps with estimation. What are the differences between a slab-only model and the spike-and-slab? In this section, we illustrate how the estimated effect size shrinks towards zero under various circumstances. We visualize the shrinkage as a function of the observed effect size, the prior on the standard deviation of effect size under the slab, the sample size, and the prior probability of the spike. We chose these parameters because the posterior distribution is fully determined by these quantities (see the online Appendix).

Figure 5.3 shows the relation between the observed effect size and the estimated effect size for the slab and for the spike-and-slab for 40 observations and 100 observations. All plots show that a smaller prior standard deviation of the slab induces some shrinkage towards zero. This effect is most obvious in the top left panel, and it makes sense, as a small prior standard deviation implies there is more prior mass near the mean of the prior, which is zero. This influence of the prior standard deviation is typically referred to as *prior shrinkage*, and it intrinsic to a Bayesian approach, but not to the spike-and-slab model. Comparing the plots between the two columns illustrates the influence of the spike; whenever the observed effect size is near zero, the estimate is shrunken towards zero in the right column but not in the left column. However, when the observed effect size is far from zero, there is little additional shrinkage to the prior shrinkage.

The shrinkage in the spike-and-slab model can be explained in the following way. Whenever the observed effect size is small, the data are well described by an effect size of zero and thus the posterior probability of the spike is substantial. As a result the marginal estimate is shrunken towards the spike's estimate, 0. In contrast, when the observed effect size is large the data are poorly described by an effect size of zero and the posterior probability of the spike is negligible. As a consequence, the estimate of the spike-and-slab is practically equivalent to the estimate of the slab. The plots in the right column of Figure 5.3 show the effect of sample size on the shrinkage. For the bottom right plot, $N = 100$, if the observed effect size is small then the estimate is still shrunken towards 0, but as the observed effect size grows the shrinkage decreases much more quickly than in the top right plot where $N = 40$. This makes sense from a signal-detection perspective. If the observed effect size is, for example, 0.3 after 40 observations, the posterior probability of the spike

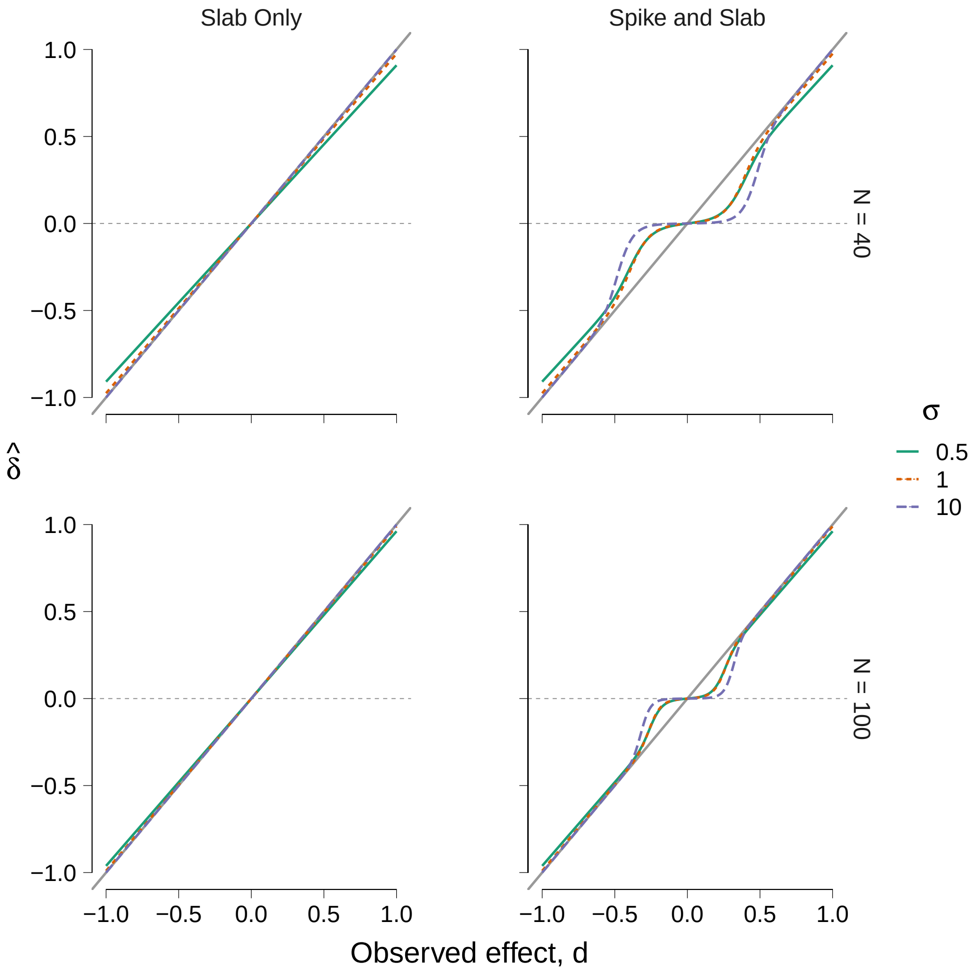


Figure 5.3: Observed effect size versus posterior mean for different model components and prior standard deviations. The left column shows inference based on the slab-only model while the right column shows inference based on the spike-and-slab model. In the top row, the sample size was 40 while in the bottom row the sample size was 100. Different lines represent different standard deviations for the prior distribution on δ. The prior probability of the spike was $^1/_2$. Inspired by Figure 5 of Rouder et al. (2018).

is substantial. However, after collecting 60 additional observations while the observed effect size remains 0.3, the posterior probability of the spike decreases as it becomes increasingly less probable that the data generating model had an effect size of zero.

Next, we explore the relationship between shrinkage and the prior probability of the spike. Figure 5.4 shows the shrinkage for various prior probabilities. The smaller the prior probability of the spike, the less the effect size is shrunken towards 0. If the prior probability is small then the spike was *a priori* implausible and less evidence is needed to make its influence negligible.

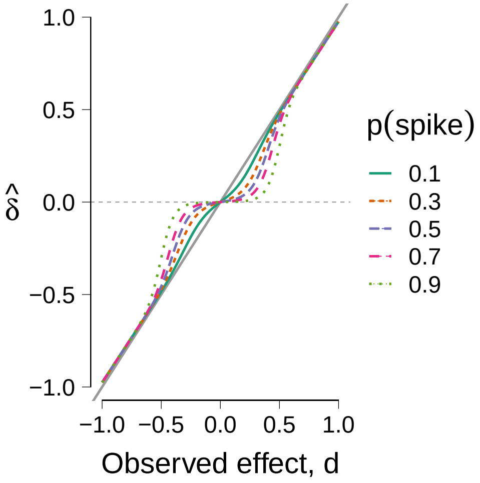


Figure 5.4: Observed effect (x-axis) versus the posterior mean of the spike-and-slab model (y-axis). The different lines represent different prior probabilities of the spike. The figure is based on 40 observations with a prior standard deviation of 1.

5.3 Empirical Example: Reanalysis of Two Minds

We now highlight how the spike-and-slab approach can be used in psychological practice by reanalyzing the results of Heycke et al. (2018), who conducted two registered replications of Rydell et al. (2006). We first briefly explain the design of the study before reanalyzing the Explicit Evaluation and Implicit Evaluation analyses with a spike-and-slab model. For a detailed description see the "Procedure" section in Heycke et al. (2018). Finally, we provide a robustness analysis.

The goal of Heycke et al. (2018) was to replicate key evidence for implicit attitude formation. In the original study, Rydell et al. (2006) reported that attitudes induced by subliminal primes manifest when they are assessed by an implicit attitude measure, and attitudes induced by supraliminal cues manifest when they are assessed by an explicit attitude measure. This finding corresponds to a perhaps surprising dissociation of implicit and explicit attitude measures. In the Heycke et al. (2018) experiments participants were briefly flashed a positive or negative prime followed by an image of a person. Next, several behavioral descriptions that were either negative or positive appeared with the image of the person (e.g., “Bob cheated during a poker game”). Afterwards, participants explicitly evaluated the target person, and performed an implicit association task (IAT). In total, data of 51 participants were analyzed. Heycke et al. (2018) could not find the dissociation between explicit and implicit attitude measures. They found that while positive descriptions resulted in a more favorable explicit evaluation than negative descriptions, positive subliminal primes did not result in more favorable IAT scores than negative subliminal primes. In contrast, both explicit and implicit attitude measures were in line with the explicit descriptions they learned during the experiment.

Explicit Evaluation In the analysis of the explicit evaluations, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that the rating of the target character is more positive if positive information is shown before negative information: $t(27) = 11.52$, $p < .001$; $\text{BF}_{10} = 1.37 \times 10^9$, $d = 2.09$, 95% HDI $[1.41, 2.79]$.[4] The magnitude of the effect is large and thus a spike-and-slab reanalysis yields practically the same results: $\hat{\delta} = 2.10$, 95% CRI: [1.74, 2.47].[5]

Implicit Evaluation In the analysis of the IAT, Heycke et al. (p. 10; 2018) conducted a paired t-test and concluded that when negative primes were presented before positive primes there was some indication that the IAT rating became more negative: $t(27) = -2.54$, $p = .017$, $\text{BF}_{10} = 2.92$, $d = -0.44$, 95% HDI $[-0.83, -0.06]$. Here, the magnitude of the effect is smaller and as a consequence the results from the spike-and-slab reanalysis are more conservative: $\hat{\delta} = -0.35$, 95% CRI: [−0.75, 0.00]. The estimate of effect size is shrunken towards 0 because the spike provides a reasonable account of the data, $\Pr(\text{spike} \mid \text{data}) = 0.25$.

Robustness analysis In the reanalyses above the prior probability of the spike was set to 0.5. One might wonder how robust or how volatile the results are to changes in the prior probability of the spike. Figure 5.5 visualizes the influence of the prior on the spike. In the left panel that shows the explicit evaluation data, the different estimates for different prior probabilities are practically identical. For this analysis, the data dominate the prior. In

[4] These are the statistics reported by Heycke et al. (2018). BF stands for Bayes factor, see also the online Appendix. HDI is short for highest density interval, a type of credible interval.

[5] The difference between the point estimate and the credible intervals is possibly caused by the difference in prior distributions for effect size. Heycke et al. (2018) use a Cauchy prior whereas we use a normal prior.

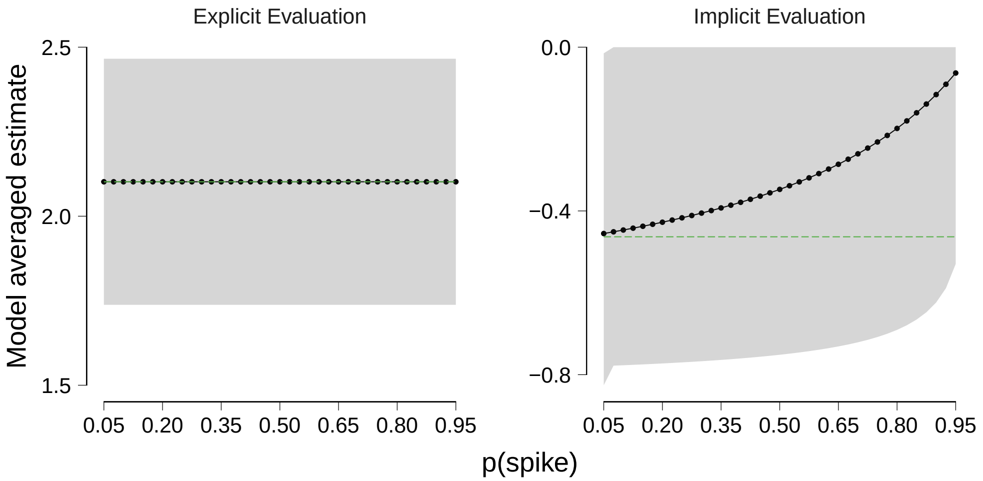


Figure 5.5: Robustness analysis that shows the prior probability of the spike (x-axis) versus spike-and-slab estimates (y-axis) for the explicit evaluation (left panel) and the implicit evaluation (right panel). Solid points show the point estimate of the spike-and-slab and the gray area represents the accompanying 95% credible interval. The green horizontal dashed line shows the estimate of the slab.

contrast, in the right panel that shows the implicit evaluation data, the prior probability of the spike has a large impact on the results. Here, the data are less informative and the prior has more influence. The adaptive shrinkage is a key feature of the spike-and-slab, that is, the amount of shrinkage depends on the posterior plausibility of the spike. Note that in the right panel the 95% credible interval becomes asymmetric as the prior and therefore also the posterior probability of the spike increases. It may appear that the credible interval is bounded by zero, however, this is a property of this particular data set. Had the observations been closer to zero then the credible interval would have also contained negative values (e.g., the posterior mass in Figure 5.2 is not zero for negative values of effect size).

5.4 Discussion

Standard estimates of effect size ignore the null hypothesis and are therefore overconfident, that is, farther away from zero than they should be. The spike-and-slab model tempers the enthusiasm that the standard estimates instill by explicitly considering the possibility that an effect is absent (Robinson, 2019; Rouder et al., 2018). The core idea dates back to Jeffreys (1939; see also Jeffreys, p. 365, 1961; Ly and Wagenmakers, 2022); nonetheless, it has been largely ignored in empirical practice, in statistical education, and in journal guidelines. We believe the spike-and-slab model is a useful statistical tool to make the interpretation of effect size estimates more robust. The spike-and-slab model optimally shrinks effect sizes with ambiguous statistical support towards zero. This data-driven statistical skepticism is appropriate regardless of whether or not researchers follow good research practices, for example, preregistering study design and analysis.

What if All Null Hypotheses Are False?

The spike-and-slab approach clashes with the popular estimation mindset, where it is argued that statistical significance should be abandoned in favor of estimation (Cumming, 2014; Cumming & Calin-Jageman, 2016; McShane et al., 2019; Valentine et al., 2015). One argument to forgo hypothesis testing is that all null hypotheses are false (Cohen, 1990; Meehl, 1978) and therefore there is no need to consider a component that states that an effect is exactly zero. The statistical counterargument is that, even if point null hypotheses are false, they are still mathematically convenient approximations to more complex hypotheses that allow mass on an interval close to zero (i.e., perinull hypotheses; Berger & Delampady, 1987; George & McCulloch, 1993; Ly & Wagenmakers, 2022). Thus, from a pragmatic perspective it is irrelevant whether or not null hypotheses are exactly true: in the spike-and-slab model, a narrow interval around zero will shrink estimates towards zero almost as much as the point null spike component will.

When Can the Spike be Ignored?

There are two scenarios in which the presence of the spike can safely be ignored. First, the spike may be deeply implausible. This happens most often in problems of pure estimation, such as when determining the relative popularity of two politicians or the proportion of Japanese cars on the streets of New York. In such cases, no value or interval needs to be singled out for special attention. Second, the data, or even data from prior studies, may provide overwhelming evidence that an effect is present, as in the reanalysis of the Explicit Evaluation data. When this happens, the results from a spike-and-slab model become virtually identical to those of a slab-only model: the inclusion of the spike offers no benefit but neither does it come with a statistical cost.

Conclusion

Standard methods for estimating effect size produce results that are overly optimistic. This tendency toward high estimates can be corrected by applying the spike-and-slab model that explicitly takes into account the possibility that the effect is absent. The spike-and-slab approach is not meant as a tool to downplay other researchers' findings that one disagrees with. Instead, it provides a more robust estimate of the size of an effect of high-quality studies whenever null and alternative hypothesis are plausible. We believe that the approach allows researchers a more nuanced interpretation of their own results taking into account the plausibility that there is no effect.

6

Default Bayes Factors for Testing the (In)equality of Several Population Variances

Testing the (in)equality of variances is an important problem in many statistical applications. We develop default Bayes factor tests to assess the (in)equality of two or more population variances, as well as a test for whether the population variances equal a specific value. The resulting test can be used to check assumptions for commonly used procedures such as the t-test or ANOVA, or test substantive hypotheses concerning variances directly. We show that our Bayes factor fulfills a number of desiderata. Researchers may have directed hypotheses such as $\sigma_1^2 > \sigma_2^2$, hey may want to extend $\mathcal{H}_0$ to have a null-region, or wish to combine hypotheses about equality with hypotheses about inequality, for example $\sigma_1^2 = \sigma_2^2 > (\sigma_3^2, \sigma_4^2)$. We extend our Bayes factor test to allow for these deviations from our proposed default and illustrate it on a number of practical examples. Our procedure is implemented in the R package *bfvartest*.

This chapter is published as: Dablander*, F., van den Bergh*, D., Wagenmakers, E.-J., & Ly, A. (in press). Default Bayes factors for testing the (in)equality of several population variances. *Bayesian Analysis.*

*These authors share first authorship.

Testing the (in)equality of variances is important in many sciences and applied contexts. In engineering, for example, researchers may want to assess whether a new, cheaper measurement instrument achieves the same precision as the gold standard (Sholts et al., 2011). In genetics and medicine, scientists are not only interested in studying the genetic effect on the mean of a quantitative trait, but also on its variance (Paré et al., 2010). In economics and archeology, ideas such as that increased economic production should reduce variability in products directly lead to statistical hypotheses on variances (Kvamme et al., 1996). In a court of law, one may be interested in reducing unwanted variability in civil damage awards and may want to compare how different interventions reduce this variability (Saks et al., 1997). In psychology, educational researchers may be interested in studying how the variance in pupil's mathematical ability changes across school grades (Aunola et al., 2004).

While there exist several classical p-value tests for assessing the (in)equality of population variances (e.g., Brown & Forsythe, 1974; Gastwirth et al., 2009; Levene, 1961), testing such hypotheses has received little attention from a Bayesian perspective. Such a perspective, however, would offer practitioners the possibility (a) to quantify evidence in favor of the null hypothesis (e.g., Morey, Romeijn, & Rouder, 2016), (b) allow one to incorporate prior knowledge (e.g., O'Hagan et al., 2006), (c) to use sequential sampling designs which in many cases is more cost-effective (e.g., than a fixed-N design, see Stefan et al., 2019), and (d) to translate substantive predictions more easily into statistical hypotheses by specifying equality and inequality constraints (e.g., Böing-Messing & Mulder, 2018; Hoijtink et al., 2008).

In light of these benefits and recent recommendations to go beyond p-value testing (Wasserstein & Lazar, 2016), we develop default Bayes factor tests (e.g., Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a; Ly et al., 2016b) for the (in)equality of several population variances. Our work is inspired by Jeffreys (1939, pp. 222-224), who developed a test for the "agreement of two standard errors". Equipped with our procedure, researchers are able to state graded evidence both for the case of testing assumptions of other tests (e.g., the equality of variances assumption in the Student's t-test), as well as testing order-constrained hypotheses on variances directly.

This paper is structured as follows. In Section 6.1, we introduce the problem setup and propose the default Bayes factor. In Section 6.2, we elaborate on the desiderata that the proposed Bayes factor adheres to. In Section 6.3, we discuss the special case with $K = 2$ groups, including directed and interval Bayes factors, compare our method to a fractional Bayes factor procedure proposed by Böing-Messing and Mulder (2018), and discuss testing all possible (in)equalities at once. We illustrate our default Bayes factor test and deviations from it on a number of practical examples in Section 6.4. We conclude in Section 6.5. All derivations and proofs can be found in the appendix.

6.1 Default Bayes Factor for K Groups

6.1.1 Notation and Problem Setup

The problem of testing the (in)equality of variances can be equivalently expressed in terms of variances σ_j^2 or precisions $\tau_j = \sigma_j^{-2}$. For the data we assume that $Y_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \tau_j^{-1})$, where $i \in [n_j]$ and $j \in [K]$ with the rectangular brackets embracing an integer denoting the set of positive integers up to and including that integer, e.g., $[K] := \{1, 2, \ldots, K-1, K\} \subset \mathbb{N}$.

As the K groups are assumed to be independent of each other, the data $y^{[K]}$ can be sufficiently summarized by the sample means $\bar{\boldsymbol{y}} = (\bar{y}_1, \ldots, \bar{y}_K)$, where $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ji}$ and the (unbiased) sample variances $\boldsymbol{s^2} = (s_1^2, \ldots, s_K^2)$, where $s_j^2 = \frac{1}{\nu_j} \sum_{i=1}^{n} (y_{ji} - \bar{y}_j)^2$ and where $\nu_j = n_j - 1$ is the degree of freedom of group j. As a convention, we denote K-dimensional vectors in bold, whereas an arrow is used to denote a $K-1$ dimensional vector, e.g., $\boldsymbol{s^2} = (\vec{s^2}, s_K^2)$. A subscript $+$ is used to denote summation over the vector's elements, e.g., $\boldsymbol{\tau}_+ = \sum_{j=1}^{K} \tau_j$, whereas $\vec{\vartheta}_+ = \sum_{j=1}^{K-1} \vartheta_j$, since $\vec{\vartheta} \in \mathbb{R}^{K-1}$.

The null hypothesis $\mathcal{H}_0$ states that all precisions are the same, while the alternative hypothesis $\mathcal{H}_1$ includes at least one inequality. Formally, we compare

$$\mathcal{H}_0 : \tau_j = \tau_k \text{ for all } j, k \in [K], \tag{6.1.1}$$

$$\mathcal{H}_1 : \tau_j \neq \tau_k \text{ for some } j \neq k \in [K], \tag{6.1.2}$$

regardless of the nuisance parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K) \in \mathbb{R}^K$. The null hypothesis restricts the K precisions to a single but unknown precision, whereas the alternative allows all precisions to vary freely. Including the means, the null model has $K+1$ free parameters, whereas the alternative model has $2K$ free parameters.

We rephrase the model comparison by generalizing the reparametrization proposed by Jeffreys (1939, pp. 222-224); see also Appendix E.1. More specifically, in the alternative model we reparametrize the K precisions $\boldsymbol{\tau}$ in terms of an average precision $\bar{\boldsymbol{\tau}} = \frac{1}{K}\boldsymbol{\tau}_+$ and $K-1$ proportions $\vec{\vartheta}$ with $\vartheta_j = \frac{\tau_j}{\boldsymbol{\tau}_+}$. Note that this reparametrization is invertible as it should be. In this parametrization the hypotheses translate into

$$\mathcal{H}_0 : \vartheta_j = \tfrac{1}{K} \text{ for all } j \in [K-1], \tag{6.1.3}$$

$$\mathcal{H}_1 : \vartheta_j \neq \tfrac{1}{K} \text{ for some } j \in [K-1], \tag{6.1.4}$$

regardless of the values of the nuisance parameter $\boldsymbol{\mu} \in \mathbb{R}^K$ and the average precision $\bar{\boldsymbol{\tau}} > 0$, which are common to both models.

From a Bayesian perspective, we assess the relative merits of $\mathcal{H}_0$ and $\mathcal{H}_1$ by virtue of how well they predict the data, that is, by their respective marginal likelihoods. The ratio of marginal likelihoods is known as the Bayes factor (Kass & Raftery, 1995), and its specification requires assigning priors to both the free parameters of the null and the alternative model. For the models being compared this implies one prior on the $2K$ free parameters of the alternative model, and another prior on the $K+1$ free parameters of the null model. To simplify matters, we mimic the nesting of the null model into the alternative model and choose $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$. The Bayes factor we propose is constructed from a right Haar prior $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ on the common parameters and from a (proper) Dirichlet prior $\pi_1(\vec{\vartheta})$ on the test-relevant parameters $\vec{\vartheta}$ with hyperparameters $\boldsymbol{u}$, where $u_j > 0$ for all $j \in [K]$.

In the remainder of this section we show that this choice of priors results in a Bayes factor that is analytic. In Section 6.2 we show that the proposed Bayes factor fulfills certain Bayesian model comparison desiderata.

6.1.2 The Proposed Bayes Factor

The choice for $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ is based on the observation that the hypotheses to be tested are invariant under (1) scalar multiplications of all the data points, and (2) location shifts of the data points of each sample/group. The nesting $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$ makes the use of the improper priors $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ permissible as a limit of proper priors with normalization constants cancelling due to their appearances in both the numerator and denominator of the Bayes factor (see also Hendriksen et al., 2021; Ly et al., 2016b; Robert, 2016). The derivations in Appendix E.2 show that with $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ on the nuisance parameters, the Bayes factor simplifies to

$$\mathrm{BF}_{10}(y^{[K]}) = \frac{\int\limits_{\Theta} \left(\int\limits_{\mathbb{R}_{>0}} \int\limits_{\mathbb{R}^K} f(y^{[K]} \mid \boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta})\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\bar{\boldsymbol{\tau}} \right) \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}}{\int\limits_{\mathbb{R}_{>0}} \int\limits_{\mathbb{R}^K} f(y^{[K]} \mid \boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta} = \frac{1}{K})\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\bar{\boldsymbol{\tau}}} \tag{6.1.5}$$

$$= \int_{\Theta} h(\boldsymbol{s^2} \mid \vec{\vartheta})\pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}, \tag{6.1.6}$$

where $\mathbb{R}_{>0}$ denotes the positive reals, $\Theta := \{\vec{\theta} \in \mathbb{R}^{K-1} \,|\, \vec{\theta}_{+} < 1\} \subset \mathbb{R}_{>0}^{K-1}$, and where we refer to $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ as the reduced likelihood, which is given by

$$h(\boldsymbol{s^2} \,|\, \vec{\vartheta}) := \Big(1 + \sum_{j=1}^{K-1} \tfrac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} \Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1 - \vec{\vartheta}_+)^{\frac{\nu_K}{2}} \Big(1 - \sum_{j=1}^{K-1} [1 - \tfrac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}, \tag{6.1.7}$$

where $\boldsymbol{\nu}_+ = \sum_{j=1}^{K} \nu_j$, and $\vec{\vartheta}_+ := \sum_{j=1}^{K-1} \vartheta_j$. Note that, for any proper prior $\pi_1(\vec{\vartheta})$, the nesting and the choice $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ leads to a measurement invariant Bayes factor, as desired. This is because $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ and therefore $\mathrm{BF}_{10}(y^{[K]}) = \mathrm{BF}_{10}(\boldsymbol{s^2})$ only depend on the data via the ratios of sums of squares $\frac{\nu_j s_j^2}{\nu_K s_K^2}$, and because each s_k^2 is invariant under location shifts within sample/group k.

The Dirichlet prior $\pi_1(\vec{\vartheta})$ on the test-relevant parameters is inspired by the form of $h(\boldsymbol{s^2} \,|\, \vec{\vartheta})$ and makes the proposed Bayes factor analytic. By definition of the integral form of the type D Lauricella function, the proposed Bayes factor is

$$\mathrm{BF}_{10}(\boldsymbol{s^2}) = \frac{\mathcal{B}(\frac{\boldsymbol{\nu}}{2} + \boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} \Big(1 + \sum_{j=1}^{K-1} \tfrac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} F_D\Big(\tfrac{\boldsymbol{\nu}_+}{2} \,;\, \tfrac{\vec{\nu}}{2} + \vec{u} \,;\, \tfrac{\boldsymbol{\nu}_+}{2} + \boldsymbol{u}_+ \,;\, \vec{1} - \tfrac{\overrightarrow{\nu s^2}}{\nu_K s_K^2}\Big), \tag{6.1.8}$$

where $\mathcal{B}(\boldsymbol{u}) = \frac{\Gamma(u_1)\cdots\Gamma(u_K)}{\Gamma(u_+)}$ is the multivariate beta function, $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^{K-1}$, $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1} s_{K-1}^2)$ is the $K-1$ vector of sums of squares, and where F_D is a type D Lauricella function which has the integral representation $F_D(a \,;\, \vec{b} \,;\, d \,;\, \vec{x}) = \frac{\Gamma(d)}{\Gamma(a)\Gamma(d-a)} \int_0^1 t^{a-1}(1-t)^{d-a-1}(1 - x_1 t)^{-b_1} \cdots (1 - x_{K-1} t)^{-b_{K-1}} \mathrm{d}t$ whenever $d > a$, which holds trivially since $u > 0$ always. Observe that, with Eq. (6.1.8) at hand, we also have an analytic marginal posterior for $\vec{\vartheta}$, namely,

$$\pi_1(\vec{\vartheta} \,|\, y^{[K]}) = \frac{\Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1 - \vec{\vartheta}_+)^{\frac{\nu_K}{2}} \Big(1 - \sum_{j=1}^{K-1} [1 - \frac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}}{\mathcal{B}(\frac{\boldsymbol{\nu}}{2} + \boldsymbol{u}) F_D\Big(\frac{\boldsymbol{\nu}_+}{2} \,;\, \frac{\vec{\nu}}{2} + \vec{u} \,;\, \frac{\boldsymbol{\nu}_+}{2} + \boldsymbol{u}_+ \,;\, \vec{1} - \frac{\overrightarrow{\nu s^2}}{\nu_K s_K^2}\Big)}. \tag{6.1.9}$$

The proposed Bayes factor can be computed from the sample variances and sample sizes directly. This makes it possible to re-evaluate the published literature without the need to have access to the raw data, as shown in Section 6.4. In the next section, we show that the proposed Bayes factor fulfills a number of desiderata.

6.2 Properties of the Proposed Bayes Factor

An important result of this paper is that our proposed Bayes factor fulfills a number of desiderata (Bayarri et al., 2012; Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a; Ly et al., 2016b). More specifically, we show that the proposed Bayes factor has the finite-sample properties of being (i) labelling invariant, (ii) (exactly) predictively matched, and (iii) information consistent. It also has the asymptotic properties of being (iv) model selection consistent and (v) limit and across-sample consistent. Information consistency requires $u_j \leq 1/2$ for $j \in [K]$ while labelling invariance requires $u_i = u_j$ for all $i, j \in [K]$, suggesting the default choice of $u_j = 1/2$ for all $j \in [K]$.[2]

6.2.1 Labelling Invariance

A Bayes factor is labelling invariant if it is independent of the arbitrary choice of which group is labelled K.

Theorem 6.2.1 (Labelling invariance)**.** *The proposed Bayes factor with $u_i = u_j$ for all $i, j \in [K]$ is labelling invariant.* ◇

Proof. See Appendix E.3.1. □

6.2.2 Predictive Matching

A Bayes factor is (exactly) predictively matched if it equals 1 for all data sets of insufficient size, that is, $\mathrm{BF}_{10}(y^{[K]}) = 1$ for all $y^{[K]}$ with $\boldsymbol{n} = (n_1, \ldots, n_K)$ smaller than the minimal sample sizes (Bayarri et al., 2012). The insufficient sizes are: (a) $n_1 = \ldots = n_K = 1$ as then $\nu_j s_j^2 = 0$ for all $j \in [K]$ regardless of the observations, and (b) $n_k = 2$ for some $k \in [K]$ and $n_j = 1$ for all $j \in [K] \setminus \{k\}$, in which case there is no other sample variance to compare s_k^2 to.

Theorem 6.2.2 (Predictive matching)**.** *A Bayes factor constructed from the pair of priors $\pi_1(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}, \vec{\vartheta}) = \pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}})\pi_1(\vec{\vartheta})$ and $\pi_0(\boldsymbol{\mu}, \bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ with $\pi_1(\vec{\vartheta})$ proper is predictively matched. This holds for our proposed Bayes factor.* ◇

Proof. See Appendix E.3.2. □

6.2.3 Information Consistency

Information consistency implies that for all data sets of sufficient size, that is, fixed $\boldsymbol{n} = (n_1, \ldots, n_K)$ with at least two indexes $j \neq k \in [K]$ such that

[2] Values $0 < u < 1/2$ would also fulfill all desiderata, but would put even more mass on large differences between the variances; we therefore use $u = 1/2$ as our default choice.

$n_j, n_k \geq 2$, the Bayes factor in favor of the alternative over the null should tend to infinity whenever it becomes abundantly clear that the null cannot hold true. This occurs in the limit $s_j^2 / s_K^2 \to 0$, that is, when the observed variance s_K^2 is of a much higher order than another sample variance s_j^2.

Theorem 6.2.3 (Information consistency)**.** *The proposed Bayes factor is information consistent if $u_j \leq 1/2$ for $j \in [K]$.* ◇

Proof. See Appendix E.3.3. □

6.2.4 Model Selection Consistency

A Bayes factor is model selection consistent if it selects the correct model as $\boldsymbol{n} \to \infty$, that is, if

$$\mathrm{BF}_{10}(Y^{[K]}, \boldsymbol{n}) \xrightarrow{\mathbb{P}} 0 \text{ if } \mathbb{P} \in \mathcal{M}_0, \text{ and } \mathrm{BF}_{01}(Y^{[K]}, \boldsymbol{n}) \xrightarrow{\mathbb{P}} 0 \text{ if } \mathbb{P} \in \mathcal{M}_1, \tag{6.2.1}$$

where $\mathbb{P}$ refers to the data generating distribution, and where $X_n \xrightarrow{\mathbb{P}} X$ denotes convergence in probability, that is, $\lim_{n \to \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$ for all $\epsilon > 0$.

To state the theorem and to allow the K sample sizes go to infinity independently of each other, we let $n_K := n$ and $n_j := c_j n$ for $c_j > 0$, $j \in [K]$, thus, $c_K = 1$ by definition. To also allow the (data-governing) variances to differ arbitrarily as well, we let γ_j be the relative size of the variance σ_j^2 with respect to σ_K^2, that is, $\sigma_j^2 := \gamma_j \sigma_K^2$ where $\gamma_j > 0$ for $j \in [K]$, thus, $\gamma_K = 1$ by definition. Note that the null hypothesis is equivalent to $\boldsymbol{\gamma} = \mathbf{1} \in \mathbb{R}^K$, whereas under the alternative there exists at least one $j \in [K]$ such that $\gamma_j \neq 1$.

Theorem 6.2.4 (Model selection consistency)**.** *The proposed Bayes factor is model selection consistent. Furthermore, let $Y_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \sigma_j^2)$ where $\sigma_j^2 = \gamma_j \sigma_K^2$ for $i \in [n_j]$, $n_j = c_j n$, and $n_K = n$ for $j \in [K]$, then as all the sample sizes tend to infinity, the Bayes factor behaves as*

$$\mathrm{BF}_{10}(\boldsymbol{s}^2, n) = C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma}) n^{\frac{1-K}{2}} \left(\tfrac{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle}{\boldsymbol{c}_+}\right)^{\frac{\boldsymbol{c}_+}{2} n} \Big(\prod_{j=1}^{K-1} \gamma_j^{-\frac{c_j}{2} n} \Big) \exp(V(n)), \tag{6.2.2}$$

where $\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle := \sum_{j=1}^{K} c_j \gamma_j$, $V(n) = \mathcal{O}_P(n^{-1/2})$ under the null and $V(n) = \mathcal{O}_P(n^{1/2})$ under the alternative, and where

$$C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma}) = \frac{(4\pi)^{\frac{K-1}{2}} \boldsymbol{c}_+^{\frac{1}{2}} \Big(\prod_{j=1}^{K-1} \gamma_j^{-u_j} \Big)}{\mathcal{B}(\boldsymbol{u}) \Big(\prod_{j=1}^{K-1} c_j^{\frac{1}{2}} \Big) (\boldsymbol{c}_+ - \sum_{j=1}^{K-1} \frac{c_j \gamma_j - 1}{\gamma_j})^{\boldsymbol{u}_+}}. \tag{6.2.3}$$

This means that under the alternative, $\mathcal{H}_1 : \gamma_j \neq 1$ for some $j \in [K-1]$, we have that

$$\log(\mathrm{BF}_{10}(\boldsymbol{s}^2, n)) = \log\left(C_0(K, \boldsymbol{c}, \boldsymbol{u} \,|\, \boldsymbol{\gamma})\right) + \tfrac{1-K}{2}\log(n) \\ + \Big(\boldsymbol{c}_+ \log\left(\tfrac{\langle \boldsymbol{c}, \boldsymbol{\gamma}\rangle}{\boldsymbol{c}_+}\right) - \sum_{j=1}^{K-1} c_j \log(\gamma_j)\Big)\frac{n}{2} + \mathcal{O}_P(n^{1/2}). \qquad (6.2.4)$$

Under the null, $\mathcal{H}_0 : \vec{\gamma} = \vec{1}$, this simplifies drastically, and the logarithm of the Bayes factor then behaves as

$$\log(\mathrm{BF}_{10}(\boldsymbol{s}^2, n)) = \tfrac{1-K}{2}\Big(\log(n) - \log(4\pi)\Big) + \tfrac{1}{2}\Big(\log(\boldsymbol{c}_+) - \sum_{j=1}^{K-1}\log(c_j)\Big) \\ - \boldsymbol{u}_+ \log(K) - \log \mathcal{B}(\boldsymbol{u}) + \mathcal{O}_P(n^{-1/2}). \qquad (6.2.5)$$

Hence, $\mathrm{BF}_{10}(\boldsymbol{s}^2, n)$ *converges relatively slowly to zero under the null compared to the exponential decay of* $\mathrm{BF}_{01}(\boldsymbol{s}^2, n)$ *under the alternative.* ◇

Proof. See Appendix E.3.4. □

Illustrating the Rate of Convergence

We illustrate the rate of convergence of our default Bayes factor by visualizing Equations (6.2.4) and (6.2.5) as a function of $K \in [2, 12]$ and $\gamma_1 \in [2, \ldots, 11]$ with $\gamma_2 = \ldots = \gamma_K = 1$ and $\sigma_K^2 = 1$. Equation (6.2.4) shows that under the alternative the asymptotic behavior of $\log(\mathrm{BF}_{10})$ is mostly linear in n. The left panel in Figure 6.1 shows the slope of this linear increase — termed the log Bayes factor growth — as a function of K and γ_1. We arrive at this slope by computing Equation (6.2.4) for a large number of n and regressing the result on n. When $\mathcal{H}_1$ is true, the rate of convergence of the Bayes factor is exponential, and so the log Bayes factor grows linearly. We visualize the slope of how the log Bayes factor grows across the number of groups, with larger values indicating more rapid exponential growth. We find that, as the number of groups increases, the log Bayes factor grows more quickly. This increase is also dependent on γ_1; for larger values, the Bayes factor grows more quickly with increasing number of groups.

The right panel in Figure 6.1 illustrates $\log(\mathrm{BF}_{01})$ as a function of the sample size per group for different number of groups K under the null hypothesis, using Equation (6.2.5). In contrast to the scenario when $\mathcal{H}_1$ is true, the rate of convergence when $\mathcal{H}_0$ is true is no longer exponential (see also Bahadur & Bickel, 2009; Jeffreys, 1961; Johnson & Rossell, 2010).

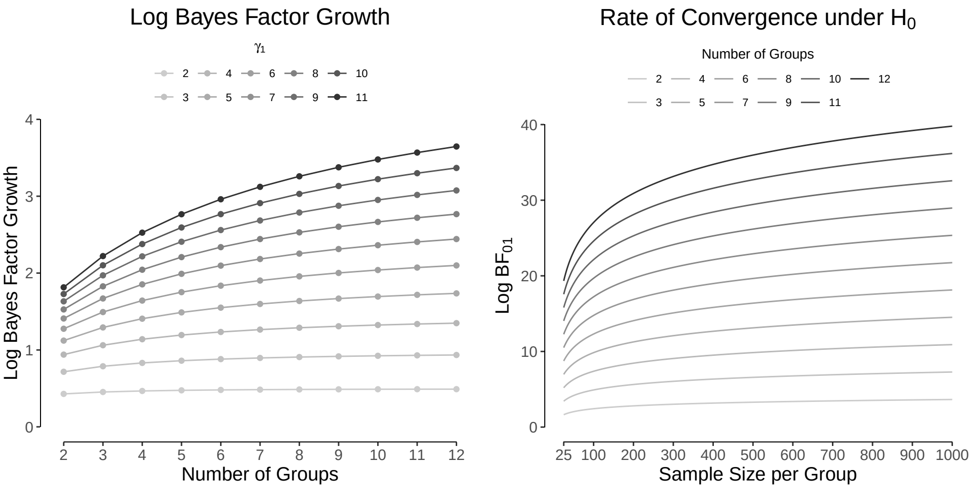


Figure 6.1: Left: Shows the rate of the linear growth of the log Bayes factor under $\mathcal{H}_1$ for increasing γ_1 and number of groups. Right: Shows how $\log(\mathrm{BF}_{01})$ grows as a function of n when $\mathcal{H}_0$ is true for different number of groups K. All Bayes factors were computed with the default value $u = 1/2$.

6.2.5 Limit and Across-Sample Consistency

A Bayes factor is limit consistent if it remains bounded as long as not all $n_j \to \infty$ for $j \in [K]$ (Ly, 2018, Ch. 6). A Bayes factor is across-sample consistent if the limit of the K-sample Bayes factor as a function of the fixed observations of the groups $i \in [K-1]$ results in a $K-1$ sample Bayes factor (Peña, 2018, Ch. 4). Note that we can consider without loss of generality the situation where the first $K-1$ samples are fixed as $n_K \to \infty$ because of labelling invariance. For the following, we assume that S_K^2 is a $\sqrt{n_K}$-consistent estimator for the data-governing variance σ_0^2 of the Kth group, which by Chebyshev's inequality is certainly the case when $Y_{Ki} \sim \mathcal{N}(\mu_K, \sigma_0^2)$.

We call the K-sample Bayes factor $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2)$ *across-sample consistent* if, as $n_K \to \infty$, it converges in probability under σ_0^{-2} to a $K-1$ Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(y^{[K-1]})$, comparing the hypotheses

$$\mathcal{H}_{0\,;\,\sigma_0^2}^{[K-1]} : \tau_j = \sigma_0^{-2} \text{ for all } j \in [K-1] \tag{6.2.6}$$

$$\mathcal{H}_{1\,;\,\sigma_0^2}^{[K-1]} : \tau_j \neq \sigma_0^{-2} \text{ for some } j \in [K-1]. \tag{6.2.7}$$

Here the null hypothesis states that the $K-1$ precisions are all equal to the known constant σ_0^{-2}, whereas the alternative states that at least one precision is unequal to σ_0^{-2}.

The theorem below implies that the proposed Bayes factor converges in probability to a lower dimensional Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ that is based on uniform priors on the nuisance parameters $\vec{\mu} \in \mathbb{R}^{K-1}$, and an inverse Dirichlet distribution on the precisions $\vec{\tau} = (\tau_1, \ldots, \tau_{K-1}) \in \mathbb{R}^{K-1}$ scaled by $1/\sigma_0^{-2}$, that is,

$$\pi_{\sigma_0^2}(\vec{\tau}\,|\,\mathcal{M}_1^{[K-1]}) = \frac{(\sigma_0^2)^{K-1}\prod_{j=1}^{K-1}(\sigma_0^2\tau_j)^{u_j-1}}{\mathcal{B}(\vec{u}, w)(1+\sigma_0^2\vec{\tau}_+)^{\vec{u}_+ + w}}, \tag{6.2.8}$$

where we wrote $w = u_K$ so the statement only involves vectors of length $K-1$. The integral representation of the multivariable generalisation of Tricomi's confluent hypergeometric function of the second kind $\mathcal{U}$, see for instance (Ng et al., 2011; Phillips, 1988), shows that the resulting $K-1$ sample Bayes factor is given by

$$\begin{aligned}\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2}) &= \frac{\int \left(\prod_{j=1}^{K-1}\tau_j^{\frac{\nu_j}{2}}\right)\exp(-\frac{1}{2}\sum_{j=1}^{K-1}\nu_j s_j^2 \tau_j)\pi_{\sigma_0^2}(\vec{\tau}\,|\,\mathcal{M}_1^{[K-1]})\mathrm{d}\vec{\tau}}{(\sigma_0^2)^{-\frac{\vec{\nu}_+}{2}}\exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})}, \\ &= \frac{\left(\prod_{j=1}^{K-1}\Gamma(\frac{\nu_j}{2}+u_j)\right)\mathcal{U}\left(\frac{\vec{\nu}}{2}+\vec{u}\,;\,\frac{\vec{\nu}_+}{2}-u_K+1\,;\,\frac{\overrightarrow{\nu s^2}}{2\sigma_0^2}\right)}{\mathcal{B}(\vec{u}, w)\exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})}, \end{aligned} \tag{6.2.9}$$

where $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1}s_{K-1}^2)$ denotes the vector of sums of squares, $(\overrightarrow{\nu s^2})_+ = \sum_{j=1}^{K-1}\nu_j s_j^2$, and $\vec{\nu}_+ := \sum_{j=1}^{K-1}\nu_j$, as before.

Theorem 6.2.5 (Limit and Across-Sample $\sqrt{n}_K$-consistency)**.** *If S_K^2 is an $\sqrt{n_K}$-consistent estimator for σ_0^2, then the Bayes factor $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2)$ is a $\sqrt{n}_K$-consistent estimator of the $K-1$-sample Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ given in Eq.* (6.2.9)*. Furthermore, if $Y_{Ki} \sim \mathcal{N}(\mu_K, \sigma_0^2)$, then $\sqrt{n}_K(S_K^2 - \sigma_0^2)$ is asymptotically normal, and consequently so is the K-sample Bayes factor, that is,*

$$\sqrt{n_K}\left(\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2) - \mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, 2\sigma_0^4\breve{T}_1^2\right) \tag{6.2.10}$$

where $\breve{T}_1$ is given by Eq. (E.3.55) *in the appendix.* ◇

Proof. See Appendix E.3.5. □

6.3 Special Cases, Deviations from the Default, and Multiple Comparisons

The comparison of $K = 2$ groups occurs frequently in practice and we discuss the Bayes factor for this special case in the following section. We also consider three modifications of the default choice in order to incorporate a subject assessment of the test-relevant parameter, and to accommodate directed tests and interval Bayes factors. Lastly, we also consider the problem of testing all possible (in)equalities, that is, the multiple comparison problem.

6.3.1 The Bayes Factor for $K = 2$ Groups

For the $K = 2$ group case, the null model of equal precisions has three parameters $(\mu_1, \mu_2, \bar{\tau})$ whereas the alternative has four $(\mu_1, \mu_2, \bar{\tau}, \vartheta)$. The comparison of interest is then between $\mathcal{H}_0 : \vartheta = \frac{1}{2}$ and $\mathcal{H}_1 : \vartheta \neq \frac{1}{2}$. In this case, the proposed Bayes factor simplifies to

$$\begin{aligned} \mathrm{BF}_{10}(\boldsymbol{s^2}) = {} & \tfrac{\mathcal{B}(\frac{\nu_1}{2}+u_1, \frac{\nu_2}{2}+u_2)}{\mathcal{B}(u_1,u_2)} \big(1 + \tfrac{\nu_1 s_1^2}{\nu_2 s_2^2}\big)^{\frac{\nu_1+\nu_2}{2}} \\ & \times {}_2F_1\big(\tfrac{\nu_1+\nu_2}{2}, \tfrac{\nu_1+2u_1}{2} \,;\, \tfrac{\nu_1+\nu_2+2(u_1+u_2)}{2} \,;\, \tfrac{\nu_2 s_2^2 - \nu_1 s_1^2}{\nu_2 s_2^2}\big), \end{aligned} \tag{6.3.1}$$

where ${}_2F_1$ refers to the Gaussian or ordinary hypergeometric function, which has the integral representation ${}_2F_1(a, b\,;\, c\,;\, z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\mathrm{d}t$, with $\mathrm{Re}(c) > \mathrm{Re}(b) > 0$ (Abramowitz & Stegun, 1972, eq. 15.3.1). Observe that across-sample consistency implies that for $Y_{2i} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_2, \sigma_0^2)$ and $n_2 \to \infty$, the two-sample Bayes factor is a $\sqrt{n}_2$-consistent estimator of the one-sample Bayes factor

$$\mathrm{BF}^{[1]}_{10\,;\,\sigma_0^2}(s_1^2) = \frac{\Gamma(\frac{\nu_1}{2} + u_1)\mathcal{U}\left(\frac{\nu_1}{2} + u_1\,;\, \frac{\nu_1}{2} - u_2 + 1\,;\, \frac{\nu_1 s_1^2}{2\sigma_0^2}\right)}{\mathcal{B}(u_1, u_2)\exp(-\frac{\nu_1 s_1^2}{2\sigma_0^2})}. \tag{6.3.2}$$

This Bayes factor compares the alternative hypothesis $\mathcal{H}^{[1]}_{1\,;\,\sigma_0^2} : \tau_1 \neq \sigma_0^{-2}$ to the null hypothesis $\mathcal{H}^{[1])}_{0\,;\,\sigma_0^2} : \tau_1 = \sigma_0^{-2}$ with σ_0^2 known. Here $\mathcal{U}(a\,;\, b\,;\, z) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-zt} t^{a-1}(1 + t)^{b-a-1}\mathrm{d}t$ is the (one-dimensional) Tricomi's confluent hypergeometric function of the second kind (Abramowitz & Stegun, 1972, Eq. 13.2.5).

6.3.2 Prior elicitation for $K = 2$ groups

For prior elicitation, it is arguably more intuitive to express the prior on the test-relevant parameter in terms of the ratio of the standard deviations, $\phi = \frac{\sigma_2}{\sigma_1} = \sqrt{\frac{\vartheta}{1-\vartheta}}$, thus, $\int_0^1 \mathrm{d}\vartheta = \int_0^\infty 2\phi(1+\phi^2)^{-2}\mathrm{d}\phi$. The prior $\vartheta \sim \mathrm{Beta}(u_1, u_2)$ underlying Eq. (6.3.1) induces a generalized beta prime distribution on ϕ with density

$$\pi(\phi\,;\, u_1, u_2) = \frac{2\phi^{2u_1-1}(1 + \phi^2)^{-(u_1+u_2)}}{\mathcal{B}(u_1, u_2)}. \tag{6.3.3}$$

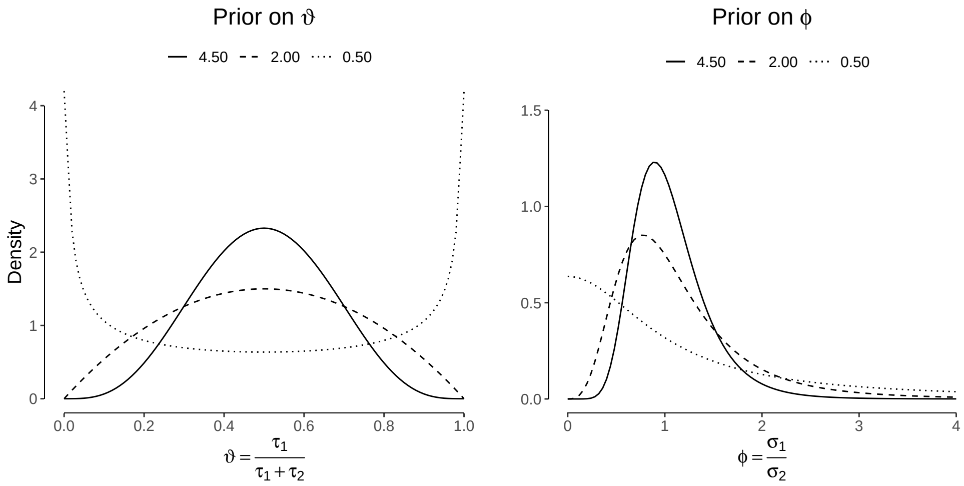


Figure 6.2: Prior on ϑ (left) and induced prior on ϕ (right) for $u := u_1 = u_2 \in \{4.50, 2.00, 0.50\}$; see Section 6.3.2 for the rationale behind these values.

Figure 6.2 visualizes the prior assigned to ϑ and ϕ for various values of $u := u_1 = u_2$. A statistician may now elicit a researcher's prior beliefs in terms of (a ratio of) standard deviations conditional on the alternative holding true. For example, if the researcher believes that the probability of one standard deviation being twice as large or twice as small as the other does not exceed 95%, then she should choose $u = 4.50$. Note that the resulting Bayes factor is not information consistent anymore. It is also interesting to note that on this scale ϕ the mth raw moment is given by $\frac{\Gamma(\frac{m}{2}+u_1)\Gamma(u_2-\frac{m}{2})}{\Gamma(u_1)\Gamma(u_2)}$. Hence, it has no finite mean whenever $u_2 \leq 1/2$. A change of variables shows that the posterior distribution in terms of ϕ is given by:

$$\pi(\phi \,|\, \boldsymbol{y}^{(2)}) = \frac{2\phi^{\nu_1+2u_1-1}(1+\phi^2)^{-(u_1+u_2)}(1+\frac{\nu_1 s_1^2}{\nu_2 s_2^2}\phi^2)^{-\frac{\nu_1+\nu_2}{2}}}{\mathcal{B}(\frac{\nu_1}{2}+u_1, \frac{\nu_2}{2}+u_2)\,{}_2F_1\left(\frac{\nu_1+\nu_2}{2}, \frac{\nu_1}{2}+u_1\,;\,\frac{\nu_1+\nu_2}{2}+u_1+u_2\,;\,1-\frac{\nu_1 s_1^2}{\nu_2 s_2^2}\right)}. \tag{6.3.4}$$

6.3.3 Interval Bayes Factors

Researchers may wish to extend the sharp null hypothesis $\vartheta = 1/2$ to include a null-region around the point null value. If the null-region overlaps with the prior under the alternative, this leads to an (inconsistent) peri-null Bayes factor (e.g., Ly & Wagenmakers, 2022; Morey & Rouder, 2011). If the null-region does not overlap with the prior under the alternative, that is, if we compare the hypotheses:

$$\mathcal{H}_0 : \phi \in [a, b] \tag{6.3.5}$$

$$\mathcal{H}_1 : \phi \notin [a, b], \tag{6.3.6}$$

then this yields a non-overlapping interval-null Bayes factor (e.g., Berger & Delampady, 1987; Rousseau, 2007). The null-region is usually informed by the problem at hand, as we will see later on an example. For a potential default approach to specify the non-overlapping interval bounds, see Appendix E.2.3.

6.3.4 Directed Bayes Factors

Researchers sometimes desire to quantify evidence in favor of hypotheses such as $\mathcal{H}_- : \sigma_1^2 > \sigma_2^2$, or $\mathcal{H}_+ : \sigma_1^2 < \sigma_2^2$. More generally, let $\mathcal{H}_r$ denote such an order-constrained or directed hypothesis. Since $\sigma_1^2 = (2\vartheta\bar{\boldsymbol{\tau}})^{-1}$ and $\sigma_2^2 = (2(1-\vartheta)\bar{\boldsymbol{\tau}})^{-1}$, we have that $\sigma_1^2 > \sigma_2^2$ implies $\vartheta < 1/2$. We therefore restrict the beta prior on ϑ accordingly in the calculation of the the marginal likelihood for $\mathcal{H}_r$ (see also Ly et al., 2016a), which can then be used to calculate directed Bayes factors.

In the more general $K > 2$ group case, we can similarly specify equality or inequality constraints by encoding them in the prior distribution on $\vec{\vartheta}$. An example of such a constrained hypotheses is given by:

$$\mathcal{H}_r : \vartheta_1 = \vartheta_2 > (\vartheta_3, \vartheta_4, \vartheta_5 = \vartheta_6) > \vartheta_7 \ \ ,$$

which incorporates two equality constraints ($\vartheta_1 = \vartheta_2$ and $\vartheta_5 = \vartheta_6$), several order constraints (e.g., $\vartheta_1 > \vartheta_3$, $\vartheta_1 > \vartheta_4$, $\vartheta_3 > \vartheta_7$, $\vartheta_4 > \vartheta_7$), and no constraints between the ϑ_3, ϑ_4, $\vartheta_5 = \vartheta_6$ (and therefore also the standard deviations and variances). Note that while this hypothesis is formulated in terms of the parameter ϑ, it has immediate implications for the precisions and thus for the standard deviations and variances. We could also directly formulate the hypotheses on the variances or standard deviations, for example, with $(\sigma_1 = \sigma_2) > \sigma_3$ implying that $(\vartheta_1 = \vartheta_2) < \vartheta_3$. This flexibility allows researchers to translate substantive predictions directly into statistical hypotheses.

We compute Bayes factors including mixed hypotheses such as $\mathcal{H}_r$ as follows. First, we introduce a new auxiliary hypothesis $\mathcal{H}_a$ which does not include order-constraints. In our example, this yields:

$$\mathcal{H}_a : \vartheta_1 = \vartheta_2, \vartheta_3, \vartheta_4, \vartheta_5 = \vartheta_6, \vartheta_7 \ \ .$$

We estimate the (auxiliary) Bayes factor BF_{ra} by dividing the proportion of samples ϑ that respect the order-constraints in $\mathcal{H}_r$ in the posterior by the proportion of samples that respect it in the prior (Klugkist et al., 2005). Separately, we then estimate the Bayes factor in favor of $\mathcal{H}_a$ over $\mathcal{H}_1$ (or $\mathcal{H}_0$)

using bridge sampling (Gronau, Sarafoglou, et al., 2017; Meng & Wong, 1996). Combining these two Bayes factors yields the desired Bayes factor in favor of $\mathcal{H}_r$ over $\mathcal{H}_1$ (or $\mathcal{H}_0$), that is, $\text{BF}_{r1} = \text{BF}_{ra} \times \text{BF}_{a1}$. The R package *bfvartest*, which is available from https://github.com/fdabl/bfvartest, implements this and all other procedures described above; see Appendix E.4 for how to use the package.

6.3.5 Comparison to a Fractional Bayes Factor

One alternative to choosing the prior based on desiderata, as done in this paper, is to use the data to inform the prior. O'Hagan (1995) proposed the *fractional* Bayes factor, which uses a fraction $b = {}^{m_0}\!/_{n}$ of the entire likelihood to construct a prior, where m_0 is the size of the minimal training sample and n is the sample size. Böing-Messing and Mulder (2018) developed a fractional Bayes factor for testing the (in)equality of several population variances. Here, we compare our proposed default Bayes factor to their fractional Bayes factor.

Since the likelihood is the same, the key difference between the two Bayes factors is in their respective prior specification. As we are concerned with hypotheses that can feature both inequality and equality constrains, we need to introduce additional notation. Let $\mathcal{H}_r$ denote a hypothesis with q_r^E equality and q_r^I inequality constraints on K population variances, such that there are $J_r = K - q_r^E$ unique variances $\vec{\sigma}_r^2 = (\sigma_1^2, \ldots, \sigma_{J_r}^2)$. Further, let K_j be the number of populations sharing the unique variance σ_j^2, and n_{j_k} be the sample size of the k^{th} population sharing the unique variance σ_j^2. Böing-Messing and Mulder (2018) use population-specific fractions given by $b_{j_k} = {}^{2}\!/_{n_{j_k}}$, where $m_0 = 2$ is the minimal training sample size for the automatic prior to be proper; it is in this sense that their Bayes factor relies on minimal prior information. They calculate the marginal likelihood for hypothesis $\mathcal{H}_r$ as:

$$p(y^{[K]} \mid \mathcal{H}_r) = \frac{\int_{\Omega_t} \int_{\mathbb{R}^K} f(y^{[K]}; \boldsymbol{\mu}, \vec{\sigma_r}^2) \pi(\boldsymbol{\mu}, \vec{\sigma_r}^2) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\vec{\sigma_r}^2}{\int_{\Omega_t^a} \int_{\mathbb{R}^K} f(y^{[K]}; \boldsymbol{\mu}, \vec{\sigma_r}^2)^{\boldsymbol{b}} \pi(\boldsymbol{\mu}, \vec{\sigma_r}^2) \mathrm{d}\boldsymbol{\mu} \mathrm{d}\vec{\sigma_r}^2} \quad , \tag{6.3.7}$$

where $\boldsymbol{b}$ is the vector of population-specific fractions, $\pi(\boldsymbol{\mu}, \vec{\sigma}_r^2) \propto \prod_{i=1}^{J_r} \sigma_i^{-2}$ is the Jeffreys prior, Ω_t specifies the region of integration depending on the inequality constraints in $\mathcal{H}_t$, and Ω_t^a is the adjusted integration region given by:

$$\Omega_t^a = \left\{ \vec{\sigma_r}^2 : \boldsymbol{R}^I [a_1 \sigma_1^2 \ldots a_{J_r} \sigma_{J_r}^2] > \vec{0} \right\} \quad , \tag{6.3.8}$$

where $\boldsymbol{R}^I$ encodes the inequality constraints among the J_r unique variances, and where $a_j = {}^{K_j}\!/_{2} \sum_{k=1}^{K_j} \left(1 - \frac{s_{j_k}^2}{n_{j_k}}\right)$. Böing-Messing and Mulder (2018) show that this setup leads to the following expression for the marginal likelihood of

$\mathcal{H}_r$:

$$p(y^{[K]} \mid \mathcal{H}_r) = \frac{\int_{\Omega_r} \prod_{j=1}^{J_r} \mathrm{IG}\left(\sigma_j^2; \frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}, \frac{\sum_{k=1}^{K_j} (n_{j_k} - 1) s_{j_k}^2}{2}\right) \mathrm{d}\sigma_j^2}{\int_{\Omega_r} \prod_{j=1}^{J_r} \mathrm{IG}\left(\frac{K_j}{\sum_{k=1}^{K_j}\left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2}\sigma_j^2; \frac{K_j}{2}, \frac{K_j}{2}\right) \mathrm{d}\sigma_j^2} \pi^{\frac{-\sum_{j=1}^{J_r}\sum_{k=1}^{K_j}(n_{j_k}-2)}{2}}$$

$$\left(\prod_{j=1}^{J_r}\prod_{k=1}^{K_j}\left(\frac{n_{j_k}}{2}\right)^{\frac{1}{2}}\right)\prod_{j=1}^{J_r}\frac{\Gamma\left(\frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}\right)\left(\sum_{k=1}^{K_j}\left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2\right)^{\frac{K_j}{2}}}{\Gamma\left(\frac{K_j}{2}\right)\left(\sum_{k=1}^{K_j}(n_{j_k} - 1) s_{j_k}^2\right)^{\frac{\sum_{k=1}^{K_j} n_{j_k} - K_j}{2}}}, \tag{6.3.9}$$

where $\mathrm{IG}(x; \alpha, \beta)$ is the density of the inverse Gamma distribution, and the ratio of the two integrals gives the probability that the constraints hold in the posterior divided by the probability that they hold in the prior. This ratio equals 1 when testing hypotheses without order-constraints, i.e., $\Omega_t^\alpha = \Omega_t$. From Equation (6.3.9) it follows that the prior distribution assigned to σ_j^2 under hypothesis $\mathcal{H}_r$ is given by:

$$\sigma_j^2 \sim \mathrm{IG}\left(\frac{K_j}{2}, \frac{\sum_{k=1}^{K_j}\left(2 - \frac{1}{n_{j_k}}\right) s_{j_k}^2}{2}\right),$$

where n_{j_k} and $s_{j_k}^2$ are the sample size and the sum of squares of the k^{th} group sharing population variance σ_j^2. Note that, in contrast to our proposed default prior, the prior for the fractional Bayes factor proposed by Böing-Messing and Mulder (2018) depends on the data. Similarly, our prior specification results in a joint distribution on $\boldsymbol{\sigma^2}$ that cannot be factorized, that is, it results in a dependent prior, where the dependency is created through the weights $\vec{\vartheta}$. The prior specification by Böing-Messing and Mulder (2018) induces a Dirichlet prior on $\vec{\vartheta}$ with $u = {}^{K_j}\!/2$ and a non-standard prior on $\bar{\tau}$ (it follows a Gamma distribution if and only if all sample sizes and sum of squares are equal). Figure 6.3 shows our default Bayes factor and the fractional Bayes factor for $K = 2$, sample sizes $n := n_1 = n_2 \in [5, \ldots, 200]$, and different values of $\phi = \{1, 1.2, 1.3, 1.4, 1.5\}$. While our proposed default Bayes factor and the fractional Bayes factor differ, they show very similar results for $u = {}^1\!/2$.

There an interesting discrepancy between the two Bayes factors when testing directed hypotheses. In case there is overwhelming evidence for the hypothesis that $\mathcal{H}_r : \sigma_1^2 > \ldots > \sigma_K^2$, the Bayes factor in favor of it over $\mathcal{H}_1 : \sigma_1^2 \neq \ldots \neq \sigma_K^2$ reaches the bound $K!$. However, in case there are the same J equalities in both

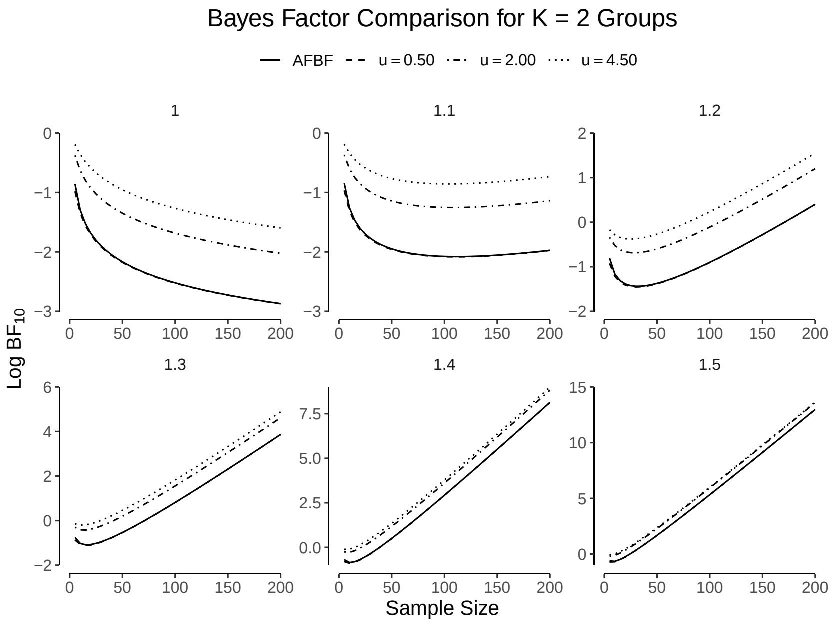


Figure 6.3: Comparison of the Bayes factor proposed by Böing-Messing and Mulder (2018) and our Bayes factor for $K = 2$ groups as a function of $n := n_1 = n_2$, prior specification $u := u_1 = u_2$, and effect size $\phi = \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$.

hypotheses, the fractional Bayes factor does not reach the bound of $(K - J)!$, while our proposed default Bayes factor does. This is because Böing-Messing and Mulder (2018) set $b_{j_k} = {}^{2}/_{n_{j_k}}$ for all groups. While this is desirable in the sense that one thus uses the same 'minimal' amount of information under each hypothesis, this results in a different shape parameter of the inverse gamma prior distribution, and the bound is therefore not reached, which can be considered a shortcoming of the fractional Bayes factor.

6.3.6 Multiple Comparisons

So far, we have focused on comparing the null hypothesis $\mathcal{H}_0$ in which all variances are equal against the alternative hypothesis $\mathcal{H}_1$ in which all variances were free to vary or against mixed hypotheses $\mathcal{H}_r$ which allow for inequalities, equalities, and order-constraints. However, researchers are sometimes also interested in assessing all possible (in)equalities. Statistically, all possible configurations of equality and inequality constraints can be uniquely represented as partitions of the groups, where any number of groups are equal if they are in the same partition. Given K groups, the number of partitions of size j is given by the Stirling numbers of the second kind, denoted $\left\{ {K \atop j} \right\}$. The total number of partitions is given by the K^{th}-Bell number, which is defined as a sum over the Stirling numbers:

$$B_K = \sum_{j=0}^{K} \left\{ {K \atop j} \right\} \, . \tag{6.3.10}$$

The Bell numbers grow quickly, with $K = 10$ already yielding $115,975$ models. This results in a multiple comparison problem, which in a Bayesian framework can be addressed by suitable adjusting the prior model odds (e.g., Jeffreys, 1961; Westfall et al., 1997). Inspired by the work on variable selection in regression (Scott & Berger, 2006, 2010), van den Bergh and Dablander (2022) recently proposed a beta-binomial prior for this problem, comparing it to a Dirichlet process prior proposed by Gopalan and Berry (1998) as well as to other methods to multiple comparison that do not require specifying a prior over all models (de Jong, 2019; Jeffreys, 1961; Westfall et al., 1997). For a small number of groups, one can directly calculate the marginal likelihood of each model and use the posterior model probabilities for inference:

$$p(\mathcal{H}_j \mid y^{[K]}) = \frac{p(y^{[K]} \mid \mathcal{H}_j)\pi(\mathcal{H}_j)}{\sum_{i=0}^{B_K} p(y^{[K]} \mid \mathcal{H}_i)\pi(\mathcal{H}_i)} = \frac{\text{BF}_{j0}\pi(\mathcal{H}_j)}{\sum_{i=0}^{B_K} \text{BF}_{i0}\pi(\mathcal{H}_i)} \, , \tag{6.3.11}$$

where B_K is the K^{th} Bell number and the prior models probabilities $\pi(\mathcal{H}_j)$ are suitable adjusted, as detailed in van den Bergh and Dablander (2022). Table 6.1 shows the results of an analysis detailed in Section 6.4.6 for a $K = 4$ group case under different model priors. For details, we refer the interested reader to van den Bergh and Dablander (2022), who also develop a stochastic search method to deal with larger K.

6.4 Practical Examples

In the following sections we apply our proposed Bayes factor test on a number of examples.

6.4.1 Sex Differences in Personality

There is a rich history of research and theory about differences in variability between men and women, going back at least to Charles Darwin (Darwin, 1871). Borkenau et al. (2013) studied whether men and women differ in the variability of personality traits. Here, we focus on peer-rated conscientiousness in Estonian women and men ($s_f^2 = 15.6$, $s_m^2 = 19.9$, $n_f = 969$, $n_m = 716$). The left panel in Figure 6.4 visualizes the raw data, and the middle panel

	Beta-binomial Prior		Dirichlet Process Prior	
Hypothesis	$\alpha = 1, \beta = 1$	$\alpha = 1, \beta = 4$	$\alpha = 1$	$\alpha = 1.817$
{Flemish, German, Estonian, Czech}	0.250 (0.446)	0.571 (0.739)	0.250 (0.368)	0.116 (0.192)
{Flemish}, {German, Czech}, {Estonian}	0.042 (0.029)	0.019 (0.007)	0.042 (0.016)	0.064 (0.034)
{Flemish, Estonian}, {German}, {Czech}	0.042 (0.005)	0.019 (0.001)	0.042 (0.003)	0.064 (0.006)
{Flemish, Czech}, {German}, {Estonian}	0.042 (0.000)	0.019 (0.000)	0.042 (0.000)	0.064 (0.000)
{Flemish}, {German, Estonian}, {Czech}	0.042 (0.083)	0.019 (0.018)	0.042 (0.053)	0.064 (0.118)
{Flemish, German}, {Estonian}, {Czech}	0.042 (0.015)	0.019 (0.004)	0.042 (0.009)	0.064 (0.023)
{Flemish}, {German}, {Estonian, Czech}	0.042 (0.018)	0.019 (0.004)	0.042 (0.015)	0.064 (0.029)
{Flemish, Estonian}, {German, Czech}	0.036 (0.030)	0.041 (0.017)	0.042 (0.014)	0.035 (0.019)
{Flemish, German}, {Estonian, Czech}	0.036 (0.060)	0.041 (0.038)	0.042 (0.056)	0.035 (0.049)
{Flemish, Czech}, {German, Estonian}	0.036 (0.004)	0.041 (0.002)	0.042 (0.003)	0.035 (0.004)
{Flemish, Estonian, Czech}, {German}	0.036 (0.005)	0.041 (0.004)	0.083 (0.009)	0.070 (0.007)
{Flemish, German, Estonian}, {Czech}	0.036 (0.061)	0.041 (0.041)	0.083 (0.105)	0.070 (0.111)
{Flemish, German, Czech}, {Estonian}	0.036 (0.003)	0.041 (0.002)	0.083 (0.005)	0.070 (0.005)
{Flemish}, {German, Estonian, Czech}	0.036 (0.211)	0.041 (0.120)	0.083 (0.339)	0.070 (0.390)
{Flemish}, {German}, {Estonian}, {Czech}	0.250 (0.029)	0.029 (0.001)	0.042 (0.003)	0.116 (0.012)

Table 6.1: Prior (and posterior) probabilites of the different hypotheses under different model priors illustrated on the example discussed in Section 6.4.6. Groups with the same population variance are put into the same set, e.g. $\sigma_1 = \sigma_2 \neq \sigma_3 = \sigma_4$ corresponds to $\{\{\sigma_1, \sigma_2\}, \{\sigma_3, \sigma_4\}\}$.

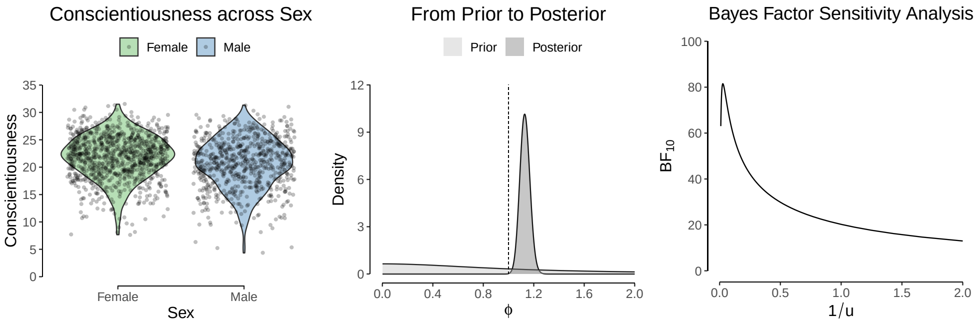


Figure 6.4: Left: Peer-rated conscientiousness of Estonian men and women. Middle: Prior and posterior of ϕ (with $u = 1/2$). Right: Bayes factor sensitivity analysis for $u \in [1/2, 100]$.

shows the prior (using $u = 1/2$) and the posterior distribution for the effect size ϕ. The default Bayes factor yields $\text{BF}_{10} = 12.98$ in favor of a difference in variance, and the right panel shows a sensitivity analysis to the specification of u in the default Bayes factor (note that the x-axis scale is $1/u$); as expected, a smaller value of u corresponds to a wider prior of ϕ under $\mathcal{H}_1$ and decreases the predictive performance of $\mathcal{H}_1$ compared to $\mathcal{H}_0$. Nevertheless, across the range of u visualized in Figure 6.4, there is strong evidence that Estonian men show larger variability in conscientiousness than Estonian women. For comparison, a frequentist analysis using Bartlett's test (Bartlett, 1937) yields $\chi^2(1) = 12.54$, $p = 0.0004$. The Vovk-Sellke bound $1/(-e \cdot p \log(p))$ (Sellke et al., 2001; Vovk, 1993) gives the maximum possible odds in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$ based on the p-value, and yields 118.11.

6.4.2 Testing Against a Single Value

Polychlorinated biphenyls (PCB), which are used in the manufacture of large electrical transformers and capacitors, are hazardous contaminants when released into the environment. Suppose that the Environmental Protection Agency is testing a new device for measuring PCB concentration (in parts per million) in fish, requiring that the instrument yields a variance of less than 0.10 (a standard deviation $\sigma_0 \leq 0.32$), thus $\phi > 1$. This suggests the use of a directed Bayes factor. Seven PCB readings on the same sample of fish are subsequently performed, yielding a sample standard deviation of $s = 0.22$ and a sample effect size of $\hat{\phi} = \frac{\sigma_0}{s} = 1.42$ (see Mendenhall & Sincich, 2016, p. 420). We compare the following hypotheses

$$\mathcal{H}_0 : \phi = 1$$
$$\mathcal{H}_+ : \phi > 1,$$

which yields $\text{BF}_{+0} = 0.51$ for the default value $u = 1/2$, a value slightly higher than for an undirected test, $\text{BF}_{10} = 0.41$. To illustrate prior elicitation, assume that the makers of the new device are highly confident, assigning 50% probability to the outcome that the new device reduces the required standard deviation at least by half. Defining $\phi = \frac{\sigma_0}{\sigma_{\text{device}}}$, this formally translates into $\pi(\phi \in [2, \infty]) = 1/2$, which is fulfilled by a (truncated) prior with $u = 2.16$. Using this prior specification results in $\text{BF}_{+0} = 0.83$.

6.4.3 Comparing Measurement Precision

In paleoanthropology, researchers study the anatomical development of modern humans. An important problem in this area is to adequately reconstruct excavated skulls. Sholts et al. (2011) compared the precision of coordinate measurements of different landmark types on human crania using a 3D laser scanner and a 3D digitizer. They reconstructed five excavated skulls and found — for landmarks of Type III, that is, the smooth part of the forehead above and between the eyebrows — an average (across skulls) standard deviation of 0.98 for the Digitizer ($n_1 = 990$) and an average standard deviation of 0.89 for the Laser ($n_2 = 990$). We define $\phi = \frac{\sigma_{\text{Digitizer}}}{\sigma_{\text{Laser}}}$ and observe that the sample effect size is 1.10. We demonstrate two tests. First, we test whether the Laser has a lower standard deviation than the Digitizer, writing:

$$\begin{aligned} \mathcal{H}_0 &: \phi = 1 \\ \mathcal{H}_+ &: \phi > 1 \ . \end{aligned}$$

The default Bayes factor in favor of $\mathcal{H}_1$ is $\text{BF}_{+0} = 4.93$ — about double the undirected Bayes factor $\text{BF}_{+0} = 2.47$ — indicating moderate evidence for the hypothesis that a 3D Laser is a more precise tool for measuring Type III landmarks on the excavated human scull compared to a 3D Digitizer. Second, in this specific scenario, a researcher might treat the Digitizer as being equally as precise as the Laser when its standard deviation differs by a maximum of 10%. She might then choose to compare the following non-overlapping hypotheses:

$$\begin{aligned} \mathcal{H}_0^{'} &: \phi \in [0.90, 1.10] \\ \mathcal{H}_+^{'} &: \phi > 1.10 \ . \end{aligned}$$

The Bayes factor with $u = 1/2$ in favor of $\mathcal{H}_0'$ is $\text{BF}_{0+}^{'} = 7.03$, indicating moderate support for the hypothesis that the Laser and the Digitizer have about equal performance. In general, we recommend researchers use the default Bayes factor unless substantive prior knowledge or particular circumstances justify a different test. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 9.16$, $p = 0.0025$, with a Vovk-Sellke bound of 24.76.

6.4.4 The "Standardization" Hypothesis in Archeology

Economic growth encourages increased specialization in the production of goods, which leads to the "standardization" hypothesis: increased production of an item would lead to it becoming more uniform. Kvamme et al. (1996) sought to test this hypothesis by studying chupa-pots, a type of earthenware produced by three different Philippine communities: the *Dangtalan*, where ceramics are primarily made for household use; the *Dalupa*, where ceramics are traded in a non-market based barter economy; and the *Paradijon*, which houses full-time pottery specialists that sell their ceramics to shopkeepers for sale to the general public. Thus, there is an increased specialization across these three communities. Kvamme et al. (1996) use circumference, height, and aperture as measures for the chupa-pots; here, we focus on the latter two. The authors test whether the standard deviations across these three groups are different, comparing:

$$\begin{aligned}\mathcal{H}_0 &: \sigma_1 = \sigma_2 = \sigma_3 \\ \mathcal{H}_1 &: \sigma_1 \neq \sigma_2 \neq \sigma_3 \ \ ,\end{aligned}$$

where σ_1, σ_2, and σ_3 correspond to the standard deviations of chupa-pots in the Dangtalan, Dalupa, and Paradijon communities, respectively. Since our Bayes factor test only requires summary statistics, we can test these hypotheses using the data from Table 4 in Kvamme et al. (1996). The authors observed $n = 55$ pots from the Dangtalan community with a standard deviation in aperture of 12.74; $n = 171$ pots from the Dalupa community with a standard deviation of 8.13; and $n = 117$ pots from the Paradijon community with a standard deviation of 5.83. Using our default prior choice of $u = 1/2$, we find overwhelming evidence for a difference in the standard deviations of the aperture measurements, $\log(\text{BF}_{10}) = 20$. Note that we can formulate a stronger statistical hypothesis based on the substantive "standardization" hypothesis, namely that the standard deviations in aperture *increase* from the Paradijon to the Dangtalan community, $\mathcal{H}_r : \sigma_1 > \sigma_2 > \sigma_3$. This yields even stronger evidence, $\log(\text{BF}_{r0}) = 21.80$, such that the Bayes factor in favor of $\mathcal{H}_r$ compared to $\mathcal{H}_1$ is very close to its theoretical maximum, $\text{BF}_{r1} = 5.98 \approx 3!$. If we were to use height instead of aperture measurements of the pots, which yield standard deviations of 9.60, 7.23, and 7.81, respectively, the evidence in favor of $\mathcal{H}_1$ and $\mathcal{H}_r$ compared to $\mathcal{H}_0$ would be much weaker, $\text{BF}_{10} = 2.27$ and $\text{BF}_{r0} = 2.87$, respectively. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 49.94$, $p < 0.00001$ with a (log) Vovk-Sellke bound of 20.75 for the aperture measurements and $\chi^2(1) = 7.18$, $p = 0.0277$ with a Vovk-Sellke bound of 3.71 for the height measurements.

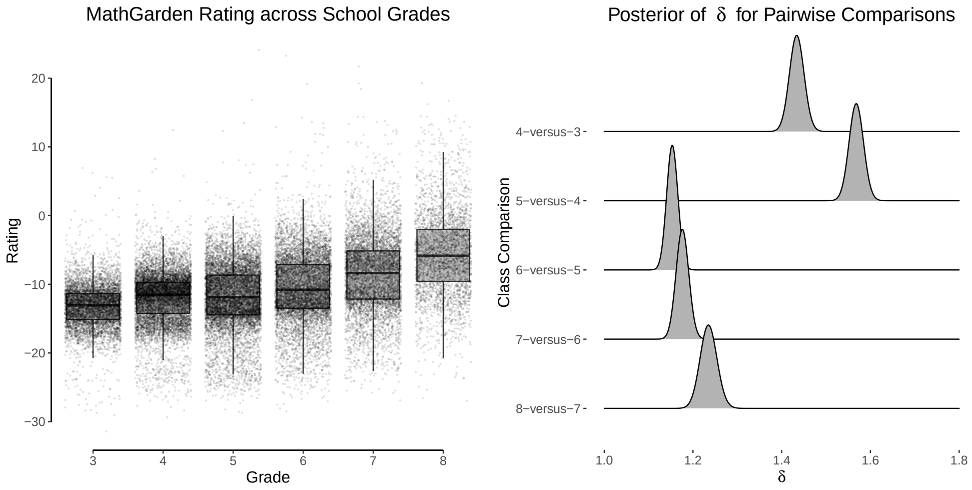


Figure 6.5: Left: Shows MathGarden rating scores across school grades. Right: Shows posterior of ϕ for pairwise consecutive class comparisons. Virtually all probability mass is assigned to $\phi > 1$, implying that, indeed, the variance increases with every school grades.

6.4.5 Increased Variability in Mathematical Ability

Aunola et al. (2004) find that the variance in mathematical ability increases across school grades. Using large-scale data from Math Garden, an online learning platform in the Netherlands (Brinkhuis et al., 2018), we assess the evidence for this hypothesis using our Bayes factor test. Math Garden assigns each pupil a rating, similar to an ELO score used in chess, and which increases if the pupil solves problems correctly. We have data from $n = 41,801$ different pupils across school grades $3 - 8$, which is visualized in the left panel of Figure 6.5. From grade 3 upwards, the standard deviations of the Math Garden ratings are $3.08, 3.69, 4.62, 4.97, 5.39$, and 5.99, for respective sample sizes of $6,410$, $9,395$, $9,160$, $7,549$, $6,007$, and $3,280$. Following Aunola et al. (2004), we wish to compare the following three hypotheses:

$$\begin{aligned}
\mathcal{H}_0 &: \sigma_i = \sigma_j \quad \forall(i,j) \\
\mathcal{H}_1 &: \sigma_i \neq \sigma_j \quad \forall(i,j) \\
\mathcal{H}_r &: \sigma_i > \sigma_j \quad \forall(i > j) \ .
\end{aligned}$$

Using the default choice $u = 1/2$, we find overwhelming support in favor of a difference in the standard deviations, $\log(\text{BF}_{10}) = 1660.53$. As is suggested by the raw data visualized in the left panel of Figure 6.5, we also find overwhelming support for an increase in variability with increased school grade, $\log(\text{BF}_{r0}) = 1667.11$. The order-constrained hypothesis again strongly outperforms the unrestricted hypothesis, yielding evidence close to its theoretical maximum, $\mathrm{BF}_{r1} = 719.69 \approx 6!$. The right panel in Figure 6.5 shows the posterior distribution of ϕ for pairwise comparisons across school grades. For comparison, Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 3366.70$, $p < 0.00001$ with a (log) Vovk-Sellke bound of 1664.07.

6.4.6 Country Differences in Conscientiousness

As our last example, we illustrate how researchers could use our default Bayes factor combined with the work by van den Bergh and Dablander (2022) to test all possible (in)equalities between variances. We utilize the data set by Borkenau et al. (2013) again, but now test whether the Czech ($s^2_C = 20, n = 714$), Estonian ($s^2_E = 17.7, n = 1685$), German ($s^2_G = 17.3, n = 303$), and Flemish ($s^2_F = 14.2, n = 291$) population differ in their variances of peer-rated conscientiousness. The posterior probability for each hypothesis under a different prior model specification can be found in Table 6.1. We find that the null hypothesis of no differences generally yields the highest posterior probability, followed by the hypothesis which states that the Flemish population variance differs from the rest. The left panels in Figure 6.6 show the posterior distributions for each variance under the full model (top) and when model-averaging across all models (bottom) using the beta-binomial($\alpha = 1, \beta = 4$), which is recommended by van den Bergh and Dablander (2022). We see that there is pronounced shrinkage towards the average variance, which is an indication that the model in which all variances are equal is strongly supported (see also Table 6.1). The right panel shows the probability that any two populations show the same variance in their peer-rated conscientiousness. We find that the German and Estonian population are most likely and the Flemish and Czech population least likely to have the same variance. This is also reflected in the unconstrained variance estimates shown in the left panel. For comparison, a Bartlett's test for $\mathcal{H}_0$ yields $\chi^2(1) = 11.51$, $p = 0.0093$ with a Vovk-Sellke bound of 8.48.

6.5 Conclusion

In this paper, we proposed a default Bayes factor test for assessing the (in)equality of several population variances and showed that it fulfills a number of desiderata for Bayesian model comparison (e.g., Bayarri et al., 2012; Consonni et al., 2018; Jeffreys, 1939; Ly et al., 2016a; Ly, 2018; Peña, 2018). In addition, we extended the Bayes factor test to cover the $K-1$-sample case, non-overlapping interval nulls, and mixed restrictions for the $K > 2$ case. The proposed procedure allows researchers to inform their statistical tests with prior knowledge. It also generalizes Jeffreys's test for the agreement of two standard errors

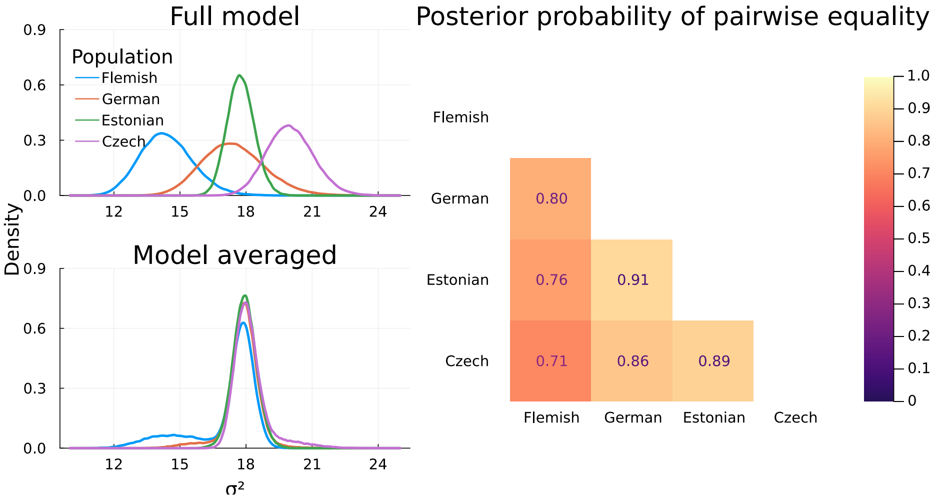


Figure 6.6: Left: Posterior means of the full model where all variances are assumed to be different (top) and posterior means when averaging across all models using a beta-binomial($\alpha = 1$, $\beta = 4$) prior (bottom). Right: Posterior probabilities for pairwise equality across all populations.

(Jeffreys, 1939, pp. 222-224); see Appendix E.1. We have also illustrated how our method — combined with specifying suitable model priors — can be used to test all possible (in)equalities between variances while adjusting for multiplicity (van den Bergh & Dablander, 2022)

A limitation of the proposed methodology is that it assumes that the data follow a Gaussian distribution, which might not always be adequate in practical applications. A potential extension would be to use a t-distributions with a small number of degrees of freedom $\nu \geq 3$, so as to better accommodate outliers, and then test whether the scales of these t-distributions differ. Another future avenue is to allow for data from the same unit, that is, allow for correlated observations or dependent groups. For the present, we believe that our work provides an elegant Bayesian complement to popular classical tests for assessing the (in)equality of several independent population variances, ready for routine applications.

7

Flexible Bayesian Multiple Comparison Adjustment Using Dirichlet Process and Beta-Binomial Model Priors

Researchers frequently wish to assess the equality or inequality of groups, but this comes with the challenge of adequately adjusting for multiple comparisons. Statistically, all possible configurations of equality and inequality constraints can be uniquely represented as partitions of the groups, where any number of groups are equal if they are in the same partition. In a Bayesian framework, one can adjust for multiple comparisons by constructing a suitable prior distribution over all possible partitions. Inspired by work on variable selection in regression, we propose a class of flexible beta-binomial priors for Bayesian multiple comparison adjustment. We compare this prior setup to the Dirichlet process prior suggested by Gopalan and Berry (1998) and multiple comparison adjustment methods that do not specify a prior over partitions directly. Our approach to multiple comparison adjustment not only allows researchers to assess all pairwise (in)equalities, but in fact all possible (in)equalities among all groups. As a consequence, the space of possible partitions grows quickly — for ten groups, there are already 115,975 possible partitions — and we set up a stochastic search algorithm to efficiently explore the space. Our method is implemented in the Julia package *EqualitySampler*, and we illustrate it on examples related to the comparison of means, variances, and proportions.

This chapter is submitted as: van den Bergh*, D., & Dablander*, F. (2022). Flexible Bayesian multiple comparison adjustment using Dirichlet process andbeta-binomial model priors. *Manuscript submitted for publication.* https://arxiv.org/abs/2208.07086.

*These authors share first authorship.

7. FLEXIBLE BAYESIAN MULTIPLE COMPARISON ADJUSTMENT USING DIRICHLET PROCESS AND BETA-BINOMIAL MODEL PRIORS

Assessing the equality or inequality of groups is a key problem in science and applied settings. If a confirmatory hypothesis is lacking, a standard approach is to first test whether all groups are equal and, if they are not, engage in multiple post-hoc comparisons. A large swathe of multiple comparisons techniques to guard against inflated false-positive errors exist in classical statistics, dating back to the work of John Tukey and others (e.g., Benjamini & Braun, 2002; Rao, 2009). From a Bayesian perspective, the problem of multiple comparisons can be addressed by changing the model prior (e.g., Berry & Hochberg, 1999; de Jong, 2019; Jeffreys, 1961; Westfall et al., 1997), an approach that has found prominent application in variable selection for regression (e.g., Scott & Berger, 2006, 2010). Here, we focus on a Bayesian multiplicity adjustment for testing the (in)equality between groups. Statistically, all possible configurations of equality and inequality constraints can be uniquely represented as partitions of the groups, where two groups are equal if they are in the same partition. In a Bayesian framework, one can adjust for multiple comparisons by constructing a suitable prior distribution over all possible partitions. This allows the researcher to explore the set of all possible equality and inequality relations among the groups while penalizing for multiple comparisons.

The first to propose a prior over all partitions to adjust for multiple hypotheses testing were, to our knowledge, Gopalan and Berry (1998), who suggested the Dirichlet process prior. Here, we propose a class of flexible beta-binomial priors for Bayesian multiple comparison adjustment, inspired by work on variable selection in regression (Scott & Berger, 2006, 2010) and explore its properties vis-à-vis previous work on multiple comparisons. More specifically, the current paper is structured as follows. In Section 7.1, we set up the problem and describe the Pólya urn scheme from which a number of priors can be derived. We characterize three such priors — the Dirichlet process, the beta-binomial, and the uniform prior — and outline our methodology in Section 7.2. In Section 7.3 we contrast the three priors, illustrate our method on a simulated example, and present a simulation study assessing the multiplicity adjustment of each prior. We also assess the method proposed by Westfall et al. (1997) and an uncorrected testing procedure based only on pairwise Bayes factors. As the space of possible partitions grows quickly — for ten groups, there are already 115,975 possible partitions — we set up a stochastic search algorithm to efficiently explore the space. Our method is implemented in Julia and available in the *EqualitySampler* package from https://github.com/vandenman/EqualitySampler. In Section 7.4, we apply our method to examples related to the comparison of proportions and variances. We conclude in Section 7.5.

7.1 Preliminary Remarks

In this section, we set up the hypothesis testing problem, discuss the relation between partitions and models, and describe Pólya's urn scheme that will unify the presentation of the priors in the following section.

7.1.1 Problem Setup

Our goal is to adjust for multiple comparisons in a flexible manner. Multiple comparisons are not a problem if we wish to compare only two hypotheses, denoted as $\mathcal{H}_0$ and $\mathcal{H}_1$. The Bayes factor quantifies how strongly we should update our prior beliefs about $\mathcal{H}_0$ relative to $\mathcal{H}_1$ after observing the data (Kass & Raftery, 1995; Ly et al., 2016a). Let group j consist of n_j observations $\vec{y}_j = \{y_{j1}, \dots, y_{jn_j}\}$ for $j \in \{1, \dots, K\}$ and $i \in \{1, \dots, n_j\}$, and let $\vec{y} = \{\vec{y}_1, \dots, \vec{y}_K\}$. The Bayes factor is given by:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \vec{y})}{p(\mathcal{H}_1 \mid \vec{y})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\vec{y} \mid \mathcal{H}_0)}{p(\vec{y} \mid \mathcal{H}_1)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{Prior odds}} \quad , \tag{7.1.1}$$

which does not depend on the number of hypotheses a researcher wishes to test.

A principled way to account for multiplicity is by adjusting the prior probability of the hypotheses (e.g., Jeffreys, 1961; Westfall et al., 1997). Suppose a researcher is interested in comparing K groups, parameterized by $\vec{\theta} = (\theta_1, \dots, \theta_K)$. She is not only interested in whether all parameters are equal ($\mathcal{H}_0$) or whether they are unequal ($\mathcal{H}_1$), but also which pairs of parameters are equal or not. In the language of classical statistics, she is interested in post-hoc comparisons. We focus on a Bayesian solution to this problem in the current paper. More specifically, going beyond classical testing, we consider the problem of assessing all possible equalities and inequalities between the groups. In general terms, the inference problem is:

$$\begin{aligned}
\rho &\sim \pi_\rho(.) \\
\vec{\theta} \mid \rho &\sim \pi_{\vec{\theta}}(.) \\
f(\vec{y}; \vec{\theta}, \rho) &= \prod_{j=1}^{K} g(\vec{y}_j; \theta_j, \phi) \quad ,
\end{aligned}$$

where ρ is a partition, ϕ is a nuisance parameter (in case it exists), and f and g are the likelihood functions. Using the posterior distribution of $\vec{\theta}$, we have that:

$$\begin{aligned} p(\mathcal{H}_0 \mid \vec{y}) &= p(\theta_1 = \theta_2 = \ldots = \theta_K \mid \vec{y}) \\ p(\mathcal{H}_1 \mid \vec{y}) &= p(\theta_1 \neq \theta_2 \neq \ldots \neq \theta_K \mid \vec{y}) \ . \end{aligned}$$

There are many more possible hypotheses, however, depending on the combination of equalities and inequalities. We can represent those as partitions, as we detail in the next section.

7.1.2 Partitions

The space of possible equality constraints for some parameter vector $\vec{\theta} = (\theta_1, \ldots, \theta_K)$ of size K is equivalent to the partitions of that vector. For example, for $K = 3$ the model that states $\theta_1 = \theta_2 \neq \theta_3$ is equivalent to the partition $\{\{\theta_1, \theta_2\}, \{\theta_3\}\}$. The space of possible models for $K = 5$ is shown in Figure 7.1. The correspondence between (in)equality constraints and partitions is useful as partitions have been studied extensively in combinatorics. Given K parameters, the number of partitions of size j is given by the Stirling numbers of the second kind, denoted $\left\{ {K \atop j} \right\}$. The total number of partitions is given by the K^{th}-Bell number, which is defined as a sum over the Stirling numbers:

$$B_K = \sum_{j=0}^{K} \left\{ {K \atop j} \right\} \ . \tag{7.1.2}$$

The Bell numbers grow very quickly, with the number of partitions for a vector $\vec{\theta}$ of size 10 being $B_{10} = 115,975$.

The Stirling numbers and Bell numbers can be generalized to the r-Stirling (Broder, 1984) and r-Bell numbers (Mezo, 2011), respectively. These generalizations help to construct conditional distributions, as we will see later. The r-Stirling numbers $\left\{ {K \atop j} \right\}_r$ give the number of partitions of size j given $K + r$ groups such that the first r parameters are all in distinct subsets. The r-Bell numbers give the total number of partitions given K parameters where the first r parameters are in distinct subsets. Specifically, we have:

$$\left\{ {K \atop j} \right\}_r = \sum_{i=0}^{K} \binom{K}{i} \left\{ {i \atop j} \right\} r^{K-i} \tag{7.1.3}$$

$$B_{K,r} = \sum_{i=0}^{K} \left\{ {K+r \atop i+r} \right\}_r \ . \tag{7.1.4}$$

Note that $\left\{ {K \atop j} \right\}_1 = \left\{ {K \atop j} \right\}$ and that $B_{K,0} = B_K$. Both the r-Stirling and r-Bell numbers are defined through recurrence relations, although explicit expressions exist which are easier to compute for large values; see Broder (1984) and Mezo (2011) for details.

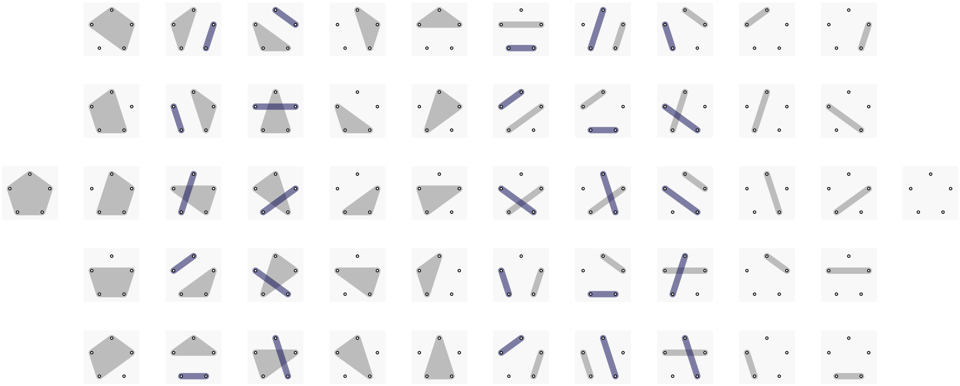


Figure 7.1: All 52 possible models given $K = 5$, represented as partitions. Circles represent individual parameters and shaded regions indicate which parameters are equal.

7.1.3 Urn Schemes

We can represent the different partitions using an urn with K different balls labeled 1 through K. For each parameter θ_j, a ball b_j is drawn from the urn with $b_j \in \{1, \ldots, K\}$. If two drawn balls are equal, $b_i = b_j$, then the two parameters are assigned to the same subset of the partition, that is, the two parameters θ_i and θ_j are equal if $b_i = b_j$. Note that different draws from an urn can represent the same partition. For example, the draws $(1, 1, 2)$ and $(3, 3, 1)$ both represent the partition $\{\{\theta_1, \theta_2\}, \{\theta_3\}\}$. The prior distributions introduced in the next sections assign probabilities to the unique partitions. Note that the prior probability of a particular draw can be obtained by dividing the probability of the corresponding partition by the total number of draws that correspond to that partition. The total number of draws that represent the same partition is given by $d!\binom{K}{d}$ where d is the number of non-empty subsets of a particular draw.

Although the urn consists of K different balls, the event of interest is whether the next ball drawn equals one of the balls already drawn — in other words, whether an equality or inequality is introduced. This event reduces the urn to a Pólya urn. All prior distributions discussed below are related to the Pólya urn. Specifically, the joint prior distribution on $(\theta_1, \ldots, \theta_K)$ is characterized by a (generalized) Pólya urn such that:

$$\theta_K \mid \theta_1, \ldots, \theta_{K-1} \sim \begin{cases} \zeta_j & \text{with probability } P_\pi \\ \theta_j^\star & \text{with probability } 1 - P_\pi \end{cases}, \tag{7.1.5}$$

where ζ_j denotes a new value for θ_K (with $\theta_1 = \zeta_1$) and $\theta_j^\star$ denotes a value equal to any previously observed value. We characterize the priors we discuss in the next section in terms of (7.1.5), which is known as a *prediction rule* (e.g., Ishwaran & James, 2001); in terms of the induced prior over partitions; and in terms of their penalty for multiplicity.

7.2 Methodology

Let $\vec{\theta}^{\star} = (\theta_1^{\star}, \ldots, \theta_r^{\star})$ denote the vector of unique population parameters out of $\vec{\theta} = (\theta_1, \ldots, \theta_K)$, $\vec{\theta}_{-j}$ the vector of parameters without parameter θ_j, and the number of repeats of $\theta_j^{\star}$ as $n_j^{\star}$. Let ρ denote a partition and $|\rho|$ its size. For example, if $\rho = \{\{\theta_1, \theta_2\}, \{\theta_3\}\}$, then $|\rho| = 2$. Similarly, for this example $\vec{\theta}^{\star} = (\theta_1^{\star}, \theta_2^{\star})$ and $n^{\star} = (2, 1)$. In the next sections, we discuss and contrast a number of priors.

7.2.1 Dirichlet Process Prior

The Dirichlet process (DP) is a distribution over distributions (Ferguson, 1973). We say that $\mathcal{G} \sim \text{DP}(\alpha, \mathcal{K})$ is distributed according to a DP if its marginal distributions are Dirichlet distributed, where α is a concentration parameter and $\mathcal{K}$ is the base distribution, which will depend on the application; for details, see for example Teh (2010). The DP can be understood as the infinite-dimensional generalization of the Dirichlet distribution, which makes it popular for mixture modeling (e.g., Rasmussen et al., 1999). Our modeling approach is similar to mixture modeling, except that we do not cluster data but parameters — a cluster corresponds to a partition. The prediction rule of the DP is given by (e.g., Blackwell & MacQueen, 1973; Ishwaran & James, 2001):

$$\theta_{j+1} \mid \theta_1, \ldots, \theta_j \sim \begin{cases} \mathcal{K} & \text{with probability } \frac{\alpha}{\alpha+j-1} \\ \text{Categorical}\left(\theta_1^{\star}, \ldots, \theta_r^{\star} \mid n_1^{\star}, \ldots, n_r^{\star}\right) & \text{else} \end{cases}, \tag{7.2.1}$$

where α is the concentration parameter and the base distribution of the DP depends on the application (see Section 7.4). In other words, we draw a new value for θ_j from $\mathcal{K}$ with probability ${}^{\alpha}/_{\alpha+j-1}$, or else set it to a previously observed value. The particular value $\theta_j^{\star}$ the parameter θ_j is set to is proportional to the number of times $\theta_j^{\star}$ was observed previously, given by $n_j^{\star}$, resulting in the well-known "rich-get-richer" property (e.g., Teh, 2010).

The Dirichlet process implies a prior distribution over partitions. The prior on the partitions ρ is:

$$\pi(\rho \mid \alpha) = \frac{\alpha^{|\rho|}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c \in \rho} \Gamma(|c|) \ , \tag{7.2.2}$$

where c is an element of ρ, and $|c|$ is its size. While the Dirichlet process features the infinite-dimensional object $\mathcal{K}$, the prior over partitions results from integrating it out. Hence the nonparametric model (in which the number of parameters is not fixed) implies a parametric model (in which the number of parameters is fixed) for the partitions (Quintana, 2006). This makes it usable for our purposes, where we have a fixed number of parameters.

The leftmost column in Figure 7.2 shows the DP prior over partitions (top) and number of inequalities (bottom) for different values of α. Intuitively, one reasonable requirement for a prior in the context of penalizing multiplicity is to be monotonically decreasing in the number of partitions, which further implies a monotonically decreasing prior probability over the number of inequalities. This is the case for $\alpha = 0.50$ (beige diamonds) as shown in the top and bottom panels, and indeed for any value $\alpha < 1$. The value suggested by Gopalan and Berry (1998) creates a symmetric prior over the partitions (yellow suns), implying that the model with no inequalities is a priori as likely as the model with all inequalities (in the $K = 5$ case, this yields $\alpha = 2.213$). The prior with $\alpha = 1$ (pink stars) results in a nonincreasing prior over the number of partitions, but in an increasing prior over the number of inequalities: the model with one inequality is more likely than the model with no inequalities.

As $\alpha \to 0$, the prior of the model with all $K - 1$ equalities $\mathcal{M}_0$ (i.e., the null model) converges to one, while as $\alpha \to \infty$, the prior of the model with $K - 1$ inequalities $\mathcal{M}_{B_K}$ (i.e., the full model) converges to one. For prior elicitation, Gopalan and Berry (1998) note that α is determined by specifying two of either $P(\mathcal{M}_0)$, $P(\mathcal{M}_{B_K})$, or their ratio, since $P(\mathcal{M}_0) = {}^{\alpha(K-1)!}/_{\prod_{j=1}^{K}(\alpha+j-1)}$ and $P(\mathcal{M}_{B_K}) = {}^{\alpha^K}/_{\prod_{j=1}^{K}(\alpha+j-1)}$; see also Table 7.1.

7.2.2 Beta-binomial Prior

The beta-binomial model prior is a popular choice for stochastic search variable selection in linear regression (George & McCulloch, 1993) and Bayesian model averaging (e.g., Hinne et al., 2020; Hoeting et al., 1999). It states that the prior probability of including j predictors out of a total of K predictors is given by:

$$\text{BB}\left(j \mid K,\ \alpha,\ \beta\right) = \binom{K}{j} \frac{\text{B}\left(j + \alpha,\ K - j + \beta\right)}{\text{B}\left(\alpha,\ \beta\right)}\ , \qquad (7.2.3)$$

where α and β are hyperparameters. The prior probability of a particular regression model is obtained by dividing by the number of ways j out of K predictors can be included: $\text{BB}\left(j \mid K,\ \alpha,\ \beta\right) / \binom{K}{j}$. The beta-binomial distribution introduces a penalty for including additional predictors and in that way introduces a correction for multiplicity (Scott & Berger, 2006, 2010).

For the multiple comparison problem discussed in this paper, we consider the number of inequality constraints and use the beta-binomial prior to in-

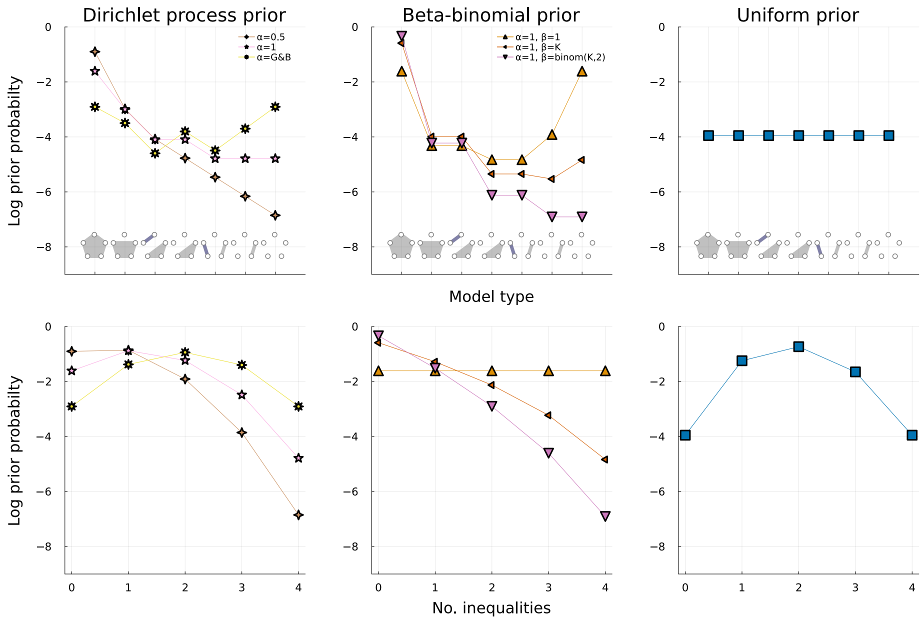


Figure 7.2: Top: Dirichlet process (left), beta-binomial (middle), and uniform prior (right) across distinct model types for $K = 5$ groups and different prior parameters. Bottom: Same but for the number of inequalities across models.

	Dirichlet process prior	Beta-binomial prior	Uniform prior
Parameters	α	$(\alpha = 1, \beta)$	✗
Prior over partitions	$\frac{\alpha^{\lvert\rho\rvert}\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{c\in\rho} \Gamma(\lvert c\rvert)$	$\binom{K-1}{\lvert\rho\rvert-1} \frac{\mathrm{B}(\lvert\rho\rvert-1+\alpha,\, K-\lvert\rho\rvert+\beta)}{\mathrm{B}(\alpha,\, \beta)\left\{{K \atop \lvert\rho\rvert}\right\}}$	$(B_K)^{-1}$
Prior monotonically decreasing	$\alpha \leq 1$	$\beta \geq K, \beta \geq \binom{K}{2}$	✗
Prior probability of null model	$\alpha(K-1)!/\prod_{j=1}^{K}(\alpha+j-1)$	$\mathrm{B}(\alpha,\, K-1+\beta)/\mathrm{B}(\alpha,\, \beta)$	$(B_K)^{-1}$
Prior probability of full model	$\alpha^K/\prod_{j=1}^{K}(\alpha+j-1)$	$\mathrm{B}(K-1+\alpha,\, \beta)/\mathrm{B}(\alpha,\, \beta)$	$(B_K)^{-1}$
Prior probability of ratio (null / full)	$\alpha(K-1)!/\alpha^K$	$\mathrm{B}(\alpha,\, K-1+\beta)/\mathrm{B}(K-1+\alpha,\, \beta)$	1

Table 7.1: Characterizations of the different priors studied in this paper. Note: $\beta \geq K$ implies a prior decreasing in terms of the number of inequalities, but not in terms of the partitions. $\beta \geq \binom{K}{2}$ implies both.

troduce a penalty for each additional inequality among the groups considered. For K groups, there can be a maximum of $K-1$ inequalities, resulting in a $\text{BB}\left(i \mid K-1,\ \alpha,\ \beta\right)$ prior distribution over the number of included inequalities i out of K groups. To see how this translates to a prior over the partitions ρ, note that there is a one-to-many correspondence between the number of inequalities i out of K groups and the resulting partitions ρ. For example, having $i=1$ inequalities with $K=3$ groups is consistent with the partitions $\{\{\theta_1,\theta_2\},\{\theta_3\}\}$, $\{\{\theta_1,\theta_3\},\{\theta_1\}\}$, and $\{\{\theta_2,\theta_3\},\{\theta_1\}\}$, all of which are of size $|\rho| = i+1$. The number of partitions of size $|\rho|$ is given, as discussed above, by the Stirling number $\left\{ {K \atop |\rho|} \right\}$. For the assignment of the prior probability, it is only the size of the partition (the number of inequalities) that counts. With these observations in hand, we arrive at the following (adjusted) beta-binomial prior distribution over partitions ρ:

$$\pi(\rho \mid K, \alpha, \beta) = \binom{K-1}{|\rho|-1} \frac{\text{B}\left(|\rho| - 1 + \alpha,\ K - |\rho| + \beta\right)}{\text{B}\left(\alpha,\ \beta\right) \left\{ {K \atop |\rho|} \right\}} \ . \tag{7.2.4}$$

The prediction rule of the beta-binomial prior is given by:

$$\theta_{j+1} \mid \theta_1, \dots, \theta_j \sim \begin{cases} \mathcal{K} & \text{with probability } P_\pi \\ \text{Categorical}\left(\theta_1^\star, \dots, \theta_r^\star \mid 1, \dots, 1\right) & \text{else }. \end{cases} , \tag{7.2.5}$$

where

$$P_\pi = \frac{\sum_{\substack{\rho \in \text{P} \\ \theta_j \notin \vec{\theta}_{-j} \subseteq \rho}} \text{BB}\left(\rho \mid K,\ \alpha,\ \beta\right)}{\sum_{\substack{\rho \in \text{P} \\ \theta_j \notin \vec{\theta}_{-j} \subseteq \rho}} \text{BB}\left(\rho \mid K,\ \alpha,\ \beta\right) + \sum_{\substack{\rho \in \text{P} \\ \theta_j \in \vec{\theta}_{-j} \subseteq \rho}} \text{BB}\left(\rho \mid K,\ \alpha,\ \beta\right)} \ , \tag{7.2.6}$$

and where P denotes the set of all possible partitions. In essence, Equation (7.2.6) takes the probability of all possible partitions where θ_j is distinct from $\vec{\theta}_{-j}$, conditional on $\vec{\theta}_{-j}$ being a subset of the considered partition. The sum over all possible partitions can be simplified using the r-Stirling numbers:

$$P_\pi = \frac{\sum_{i=1}^{K} \text{BB}\left(i \mid K,\ \alpha,\ \beta\right) \left\{ {K-j+r+1 \atop i} \right\}_{r+1}}{r \sum_{j=i}^{K} \text{BB}\left(i \mid K,\ \alpha,\ \beta\right) \left\{ {K-j+r \atop i} \right\}_{r} + \sum_{i=1}^{K} \text{BB}\left(i \mid K,\ \alpha,\ \beta\right) \left\{ {K-j+r+1 \atop i} \right\}_{r+1}} \ , \tag{7.2.7}$$

where r is number of unique parameters in $\vec{\theta}$, that is, the size of the partition.

The beta-binomial prior on the partitions and the induced prior on the number of inequalities are shown for different parameterizations in the middle column in Figure 7.2. For $\alpha = \beta = 1$, the beta-binomial distribution over the partitions has a characteristic U-shape (orange triangles). This prior specification in turn implies a uniform prior on the number of inequalities. We follow M. A. Wilson et al. (2010) who, in the context of regression, suggested to set $\alpha = 1$ as a default so that the distribution over model size (here the number of inequalities) is nonincreasing, and to scale $\beta = \lambda K$ with the number of groups to force the prior to be monotonically decreasing, with a default of $\lambda = 1$ (M. A. Wilson et al., 2010). This is illustrated as the red line (leftward pointing triangles) in Figure 7.2 using $\beta = 5$. In the multiple comparison case, we additionally investigate $\beta = \binom{K}{2}$, which implies that the prior on the number of inequalities of individual models is nonincreasing, see Appendix F.2. The purple line (upside-down triangles) in Figure 7.2 shows a decreasing prior for $\beta = \binom{5}{2} = 10$. This prior assigns the least mass to models with an increasing number of inequalities compared to all others beta-binomial priors.

Figure 7.2 shows that the DP prior makes a distinction that the beta-binomial is, by design, not making: while the beta-binomial prior assigns the same prior mass to partitions with the same number of (in)equalities, the DP prior assigns more mass to the partition with the larger cluster. For example, the beta-binomial does not distinguish between $\{\{\theta_1, \theta_2, \theta_3\}, \{\theta_4\}, \{\theta_5\}\}$ and $\{\{\theta_1, \theta_2\}, \{\theta_3, \theta_4\}, \{\theta_5\}\}$, while the DP assigns more mass to the former (see Figure 7.2). We return to this distinction in the discussion.

Lastly, note that for the beta-binomial prior we have that $P(\mathcal{M}_0) = {}^{\mathrm{B}(\alpha,\; K-1+\beta)}/_{\mathrm{B}(\alpha,\; \beta)}$ and $P(\mathcal{M}_{B_K}) = {}^{\mathrm{B}(K-1+\alpha,\; \beta)}/_{\mathrm{B}(\alpha,\; \beta)}$. Fixing $\alpha = 1$, we have that as $\beta \to \infty$, the prior of the model with all $K - 1$ equalities $\mathcal{M}_0$ converges to one, while as $\beta \to 0$, the prior of the model with $K - 1$ inequalities $\mathcal{M}_{B_K}$ converges to one; see also Table 7.1. As with the Dirichlet process prior discussed above, one can use these relations in prior elicitation.

7.2.3 Uniform Prior

For completeness, we give a prior that is uniform over the space of partitions. The probability mass function is straightforward. All valid configurations of size K have probability ${}^{1}/_{B_K}$. The prediction rule of the uniform prior is given by:

$$\theta_{j+1} \mid \theta_1, \ldots, \theta_j \sim \begin{cases} \mathcal{K} & \text{with probability } P_{\pi_U} \\ \text{Categorical}\left(\theta_1^\star, \ldots, \theta_r^\star \mid 1, \ldots, 1\right) & \text{else} \end{cases}, \tag{7.2.8}$$

where

$$P_{\pi_U} = \frac{B_{K-j+1,\, r+1}}{B_{K-j+1,\, r+1} + r B_{K-j+1,\, r}} \tag{7.2.9}$$

Here, $B_{K-j+1,\, r+1}$ counts the number of models where $\theta_{j+1} \notin (\theta_1^\star, \ldots, \theta_r^\star)$ conditional on $\theta_1, \ldots, \theta_j$ being assigned to r distinct subsets. Complementarily,

$B_{K-j+1,r}$ counts the number of models where $\theta_{j+1} \in (\theta_1^\star, \dots, \theta_r^\star)$ conditional on $\theta_1, \dots, \theta_j$ being assigned to r distinct subsets, which is multiplied by r as there are r subsets that θ_{j+1} could be assigned to. Under this uniform prior, all partitions ρ are equally likely, as can be seen in the top right panel in Figure 7.2. Note that this uniform prior induces a non-uniform prior on the number of inequalities, as shown in the bottom right panel.

7.2.4 Posterior Model Consistency

Model selection consistency is a key desiderata that a good Bayes factor should fulfill (e.g., Bayarri et al., 2012; Consonni et al., 2018; Ly et al., 2016a). In the situation of multiple models, the notion of pairwise model selection consistency needs to be extended. This extension is referred to as posterior model selection consistency. Posterior model consistency in a model class $\mathfrak{M}$ is the convergence to one, in probability, of the posterior probabilities to the true model (e.g., Casella et al., 2009; Moreno et al., 2015). Let $\mathcal{M}_j \in \mathfrak{M}$ be the model that instantiates the hypothesis $\mathcal{H}_j$ that specifies the (in)equalities among K groups. The posterior probability of $\mathcal{M}_j$ is given by:

$$p(\mathcal{M}_j \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}_j)\pi(\mathcal{M}_j)}{\sum_{i=0}^{B_K} p(\mathcal{D} \mid \mathcal{M}_i)\pi(\mathcal{M}_i)} = \frac{\text{BF}_{j0}\pi(\mathcal{M}_j)}{\sum_{i=0}^{B_K} \text{BF}_{i0}\pi(\mathcal{M}_i)} \,. \tag{7.2.10}$$

It follows that if the Bayes factor is model selection consistent, posterior model consistency holds (see also Moreno et al., 2015, Theorem 1) — unless the prior assigns zero mass to the true model. This is not the case for any of the priors discussed above, and hence whether posterior model consistency holds depends solely on the priors on the parameters within models.

7.2.5 Stochastic Search Method

When the number of groups is small and the computation of Bayes factors is swift, one can directly compute the Bayes factors for all hypotheses. Using the priors we outlined above, one can then obtain posterior distributions over hypotheses that incorporate the desired multiplicity adjustment. The number of (in)equalities grows extremely quickly with the number of groups, however, and for larger number of groups one must rely on stochastic search methods. Moreover, while directly computing the Bayes factors results in posterior distributions over hypotheses, it does not yield posterior distributions over parameters. We therefore set up a stochastic search method that yields both, allowing researchers to incorporate uncertainty across hypotheses through model averaging (e.g., Hinne et al., 2020; Hoeting et al., 1999).

Our method is implemented in the programming language Julia (Bezanson et al., 2017). First, we implemented the prior distributions in Julia. Next, we used the library *Turing.jl*, which is designed for general-purpose probabilistic programming (Ge et al., 2018). Turing enabled us to directly reuse the distributions defined in Julia code and also provided a multitude of options for composing different MCMC samplers. We set up a Gibbs sampler that explored the posterior space in two steps. The first step used Turing's built-in Hamiltonian Monte Carlo methods for sampling from the posterior distributions of the continuous parameters. In all models discussed here, all parameters are continuous except for the partitions. The second step used a custom Gibbs algorithm for sampling from the posterior distribution over partitions. The partitions were represented as a vector of integers denoted $\vec{\gamma}$ that indicate partition membership. By partition membership, we mean that two parameters θ_i and θ_j are in the same partition if and only if $\gamma_i = \gamma_j$. For example, $\{\{\theta_1\}, \{\theta_2, \theta_3\}\}$ could be represented by $(1, 2, 2)$ but also by $(3, 1, 1)$. We first explain the remainder of the sampling scheme and motivate the duplicate representations in the next paragraph. The number of possible duplicate representations in $\vec{\gamma}$ for one partition is straightforward to compute, and the prior over $\vec{\gamma}$ is obtained by taking the prior over the partitions and dividing uniformly over duplicate representations. Next, we sample each element of $\vec{\gamma}$ conditional on the other elements. Since the partition membership is discrete, we enumerate all possible values and draw from the resulting categorical distribution. Sampling individual elements of $\vec{\gamma}$ from the conditional distributions rather than the joint distribution reduces the complexity from $\mathcal{O}(B_K)$ to $\mathcal{O}(K^2)$.

Although the duplicate representations of $\vec{\gamma}$ for one partition introduce some additional computational cost, they facilitate exploration of the posterior space. For example, if we had used a one-to-one mapping from partitions to $\vec{\gamma}$, then updating the first membership in $(1, 2, 2)$ to $(2, 2, 2)$ would not be a valid configuration, as this should be represented by $(1, 1, 1)$. However, a transition from $(1, 2, 2)$ to $(1, 1, 1)$ requires updating two parameters and is therefore less likely to occur. Nevertheless, on the level of partitions, it makes sense to propose a move from $\{\{\theta_1\}, \{\theta_2, \theta_3\}\}$ to $\{\{\theta_1, \theta_2, \theta_3\}\}$.

7.3 Investigating Multiplicity Adjustment

In this section, we investigate the differences between the above priors in more detail and compare them to the method proposed by Westfall et al. (1997) and an uncorrected approach using pairwise Bayes factors. In Section 7.3.1, we use a small simulation study to illustrate the implications of multiplicity adjustment. In Section 7.3.2, we present the results of a more extensive simulation study.

7.3.1 Illustrating Multiplicity Adjustment

Here we illustrate the different multiplicity penalties that the different priors impose using a small simulation study. We simulate data from a one-way ANOVA model and analyze it using the specification by Rouder et al. (2012). The ANOVA model extended with a prior over partitions is given by:

$$\begin{aligned}
Y_{ij} &\sim \mathcal{N}\left(\mu + \sigma\theta_j, 1\right) \\
\mu &\propto 1 \\
\sigma^2 &\propto 1/\sigma^2 \\
g &\sim \mathcal{IG}\left(1/2, 1/2\right) \\
\vec{\theta}^u &\sim \mathcal{N}_{K-1}\left(0, g\right) \\
\vec{\theta}^c &\leftarrow \mathbf{Q}\vec{\theta}^u \\
\theta_j &\leftarrow \text{mean of elements of } \theta^c \text{ in the same partition} \\
\rho &\sim \pi_\rho(.) \ .
\end{aligned} \tag{7.3.1}$$

The data follow a Gaussian distribution with a grand mean μ and a group-specific offset θ_j. The offsets sum to zero to avoid identification constraints. This is achieved by projecting $\vec{\theta}^u$ from a $K - 1$ dimensional space onto a K dimensional space using the matrix $\mathbf{Q}$, which consists of the first $K - 1$ columns of an eigendecomposition of a degenerate covariance matrix as defined in Rouder et al. (2012).[2] Next, the elements of $\vec{\theta}^c$ within the same partition are averaged to obtain θ_j. The unconstrained offsets $\vec{\theta}^u$ are assigned a g prior where g itself is assigned an inverse gamma prior with shape and scale equal to $1/2$ (Liang et al., 2008). Note that the model reduces to the approach of Rouder et al. (2012) whenever the partition indicates that all elements are distinct.

We simulated from the null model, which assumes that all the groups are equal, and from the full model, which assumes that all groups are unequal, drawing 100 observations per group and varying the number of groups $K \in [2, 3, \ldots, 10]$, repeating each combination 100 times. In the full model, the means were of increasing size with successive differences of 0.20. For the analysis we considered six priors: the Dirichlet process prior with $\alpha \in \{0.50, 1\}$ and α set adaptively to have equal prior mass assigned to the model with all equalities and the model with all inequalities (i.e., $p(\mathcal{H}_0) = p(\mathcal{H}_1)$), as done by Gopalan and Berry (1998); the beta-binomial prior with $\alpha = 1$ and $\beta \in \{1, K, \binom{K}{2}\}$; and the uniform prior. We also included the prior adjustment method proposed by Westfall et al. (1997) and an uncorrected method using

[2] Note that this projection is not unique. It can also be achieved with, for example, a QR decomposition, as recommended by the Stan Development Team (2022).

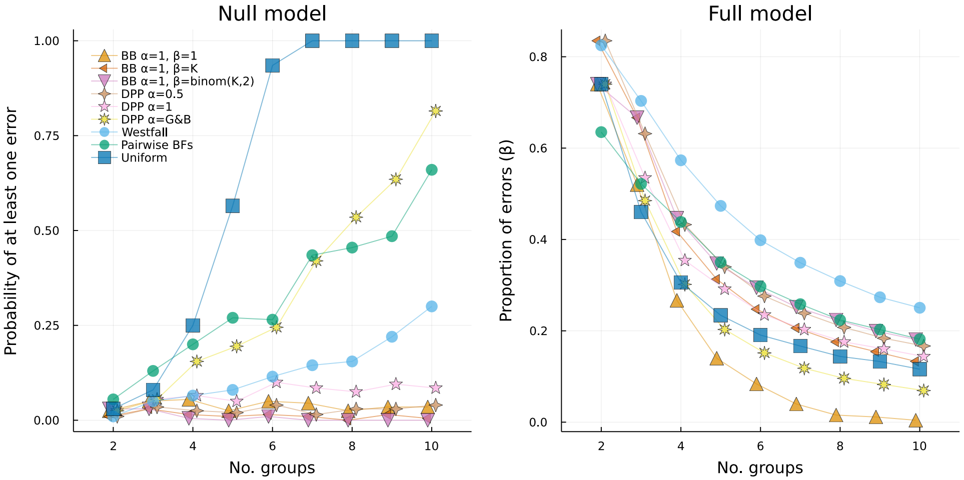


Figure 7.3: Left: Probability of making at least one false claim about a difference between two groups when there is none. Right: Proportion of falsely claiming no difference between two groups when there is one.

pairwise Bayes factors. We used our methodology as described in Section 7.2.5, drawing 12,000 MCMC samples and discarding the first 2,000 as a burn-in.

To assess how well the respective priors adjust for multiplicity, we calculated how frequently the posterior probability that any two groups differ is larger than 0.50, using the null model as data-generating model. Similarly, to assess how well the respective priors are capable of detecting true differences, we calculated how frequently the posterior probability that any two groups *do not* differ is larger than 0.50, using the full model as data-generating model.

The left panel in Figure 7.3 shows that using a uniform prior (blue squares) very quickly leads to false positives as the number of groups increases. This is not surprising: the uniform prior assigns each model the same prior mass, hence diminishing the plausibility assigned to $\mathcal{H}_0$ dramatically as K increases, thus increasing the probability of an error. The Dirichlet process prior which assigns equal mass to the full and the null model (yellow suns), as suggested by Gopalan and Berry (1998), performs better than the uniform prior but still does not provide adequate error control. It performs roughly as poorly as the method which simply computes pairwise Bayesian t-tests (green circles). The correction proposed by Westfall et al. (1997) performs much better (light blue circles) but still leads to a relatively high probability of making at least one error as the number of groups increases. The DP prior with $\alpha = 1$ (pink stars) performs better, with the DP prior with $\alpha = 0.50$ (beige diamonds) and the set of beta-binomial priors providing good error control.

The right panel in Figure 7.3 shows that the beta-binomial prior with $\alpha =$

$\beta = 1$ leads to the lowest proportion of falsely claiming no difference between two groups, followed by the Dirichlet process prior for which $p(\mathcal{H}_0) = p(\mathcal{H}_1)$ and the uniform prior. The method proposed by Westfall et al. (1997) performs worst, followed by the beta-binomial prior with $\alpha = 1$ and $\beta = \binom{K}{2}$ and the DP prior with $\alpha = 0.50$. The performance of the uncorrected pairwise Bayes factor approach is somewhere in the middle. Note that all approaches perform better as the group size increases, but this is due to our simulation design: each additional group exhibits a mean larger than the previous one by 0.20 and adds n more observations, which makes falsely claiming no difference less likely with an increasing number of groups. Instead of looking at absolute error, we therefore focus on the relative ordering of the priors. Overall, we conclude that not adjusting for multiple comparisons — either by using a uniform prior or by using pairwise Bayes factors — naturally leads to the worst performance and that the method by Westfall et al. (1997) is overly conservative and does not provide adequate error control with an increasing number of groups. In the next section, we report on a more extensive simulation study to further disentangle the differences between the multiple comparison methods.

7.3.2 Simulation Study

In the previous section, we illustrated the importance of adjusting the prior model probabilities in reducing the familywise error rate when all groups are equal. Here we explore the multiplicity adjustment of the different methods in a more exhaustive simulation study. We used the same ANOVA model as in the previous section and varied the total number of groups $K \in \{5, 9\}$ and the sample size per group $n \in \{50, 100, 250, 500\}$. In addition, we varied the true number of equalities to be $\{0\%, 25\%, 50\%, 75\%, 100\%\}$. For $K = 5$, there are 4 possible (in)equalities which resulted in models that have either 0, 1, 2, 3, or 4 equalities. For $K = 9$, there are 8 possible (in)equalities, resulting in 0, 2, 4, 6, or 8 equalities in the true model. Given the number of equalities, we sampled a particular partition uniformly from all possible partitions with that amount of equalities and used this model to simulate data. Each unique combination was repeated 100 times and each generated data set was analyzed with the same prior specifications as above. We assessed the familywise error control as well as statistical power. The results for K = 5 and K = 9 were similar. Therefore, we focus on the $K = 5$ in the main text and discuss the $K = 9$ case in Appendix F.3.

Note that the hierarchical approach has an additional source of α error in contrast to pairwise comparisons when there are more than 0 inequalities because it imposes transitivity. For example, imagine that the true model postulates that $\theta_1 = \theta_2 = \theta_3 \neq \theta_4$. However, the sample means are (by random sampling) $\bar{x}_1 = 0.1, \bar{x}_2 = 0.2, \bar{x}_3 = 0.3, \bar{x}_4 = 0.35$. The hierarchical approach would find that $\theta_3 = \theta_4$, but not that $\theta_1 = \theta_3$ since that also implies $\theta_1 = \theta_4$. Therefore, the model $\theta_1 = \theta_2 = \theta_3 \neq \theta_4$ and even the equality $\theta_1 = \theta_2$ are not retrieved. In contrast, the pairwise methods violate transitivity as they only look at two pairs at the time and will happily suggest that $\theta_1 = \theta_2$, $\theta_2 = \theta_3$, and $\theta_3 = \theta_4$ while simultaneously suggesting that $\theta_1 \neq \theta_4$.

Familywise Error Rate

Figure 7.4 shows the probability of at least one error for different methods across the number of *inequalities* in the true model and sample sizes. The top left panel shows that the uniform prior (blue squares), the pairwise Bayes factors (green circles), the Dirichlet process prior with ($p(\mathcal{H}_0) = p(\mathcal{H}_1)$) (yellow stars), and the method proposed by Westfall et al. (1997) (light blue circles) perform worst and that the other Dirichlet process and beta-binomial priors provide adequate error control. This mirrors the results above, which is natural since this part of the simulation is a special case for $K = 5$. Increasing the number of inequalities to 1 (top right) and 2 (bottom left), we find that the pairwise Bayes factors, the method by Westfall et al. (1997), and the uniform improve in performance. This is likely due to the fact that, with more inequalities, there are simply less opportunities to incorrectly claim that two population means are different. In contrast, the performance of the other methods decreases when there is at least one inequality; it is difficult to disentangle a trend with increasing inequalities.

The rightmost panel in Figure 7.4 shows the results averaged over the number of inequalities in the true model. We find that the method by Westfall et al. (1997) shows the strongest familywise error control, closely followed by the the beta-binomial priors with $\beta = K$ and $\beta = \binom{K}{2}$ and the DP prior with $\alpha = 0.50$. The pairwise Bayes factors perform similar to the Dirichlet process prior with $\alpha = 1$, with the beta-binomial prior with $\beta = 1$, the symmetric DP prior, and the uniform prior performing worst. The differences between the methods become less pronounced with increasing sample size since the data starts to dominate the prior.

Statistical Power

Figure 7.5 shows the proportion of falsely claiming a difference between two groups when there is none for different methods across the number of *equalities* in the true model and sample sizes. The top left panel shows that the beta-binomial prior with $\beta = 1$ performs best and the method proposed by Westfall et al. (1997) performs worst, again mirroring the results of the small simulation study above. Increasing the number of equalities in the true model, we find that the performance of virtually all methods decreases except for the

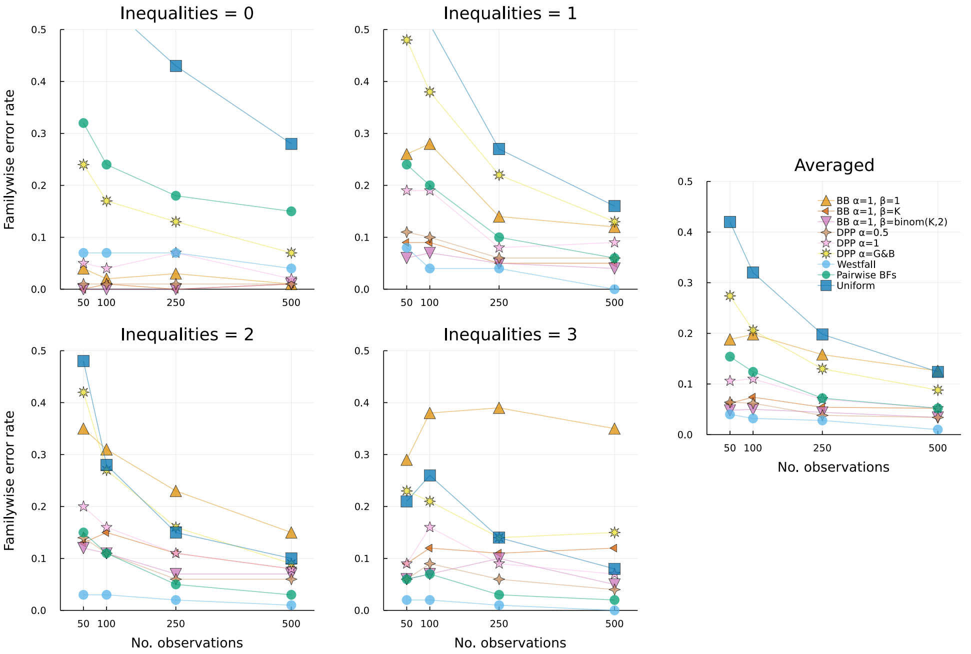


Figure 7.4: Familywise error rate across priors and sample sizes under a model with 0 (top left), 1 (top right), 2 (bottom left), and 3 (bottom right) true inequalities for $K = 5$ groups. The rightmost panel shows the average familywise error rate across inequalities.

uniform prior, which shows a slight increase, especially for large sample sizes. This overall decrease in performance is likely due to the fact that the average pairwise difference between groups *decreases* with the number of equalities. To illustrate, note that the model with no equalities for $K = 4$ groups has population means $\vec{\mu} = \{-0.30, -0.10, 0.10, 0.30\}$, which yields pairwise differences $[0.20, 0.20, 0.20, 0.40, 0.40, 0.60]$ with an average of 0.33. In contrast, including one equality results in $\vec{\mu} = \{-0.25, -0.05, 0.15, 0.15\}$[3], yielding pairwise differences of $[0.20, 0.20, 0.20, 0.40, 0.40]$ with an average of 0.28.

The rightmost panel in Figure 7.5 shows the results averaged over the number of equalities in the true model. We find that the method by Westfall et al. (1997) is highly conservative, trading off the strong familywise error control with an increase in the proportion of false negatives. Similarly, the priors that performed worst with respect to familywise error control — the uniform, symmetric DP, and beta-binomial prior with $\beta = 1$ — perform best here. The other DP and beta-binomial priors as well as the pairwise Bayes factors are

[3] This is due to the sum-to-zero constraint and the constraint that all successive unequal groups have a difference of 0.20.

somewhere in between those two extremes. Note that again the differences between the methods become less pronounced with increasing sample size.

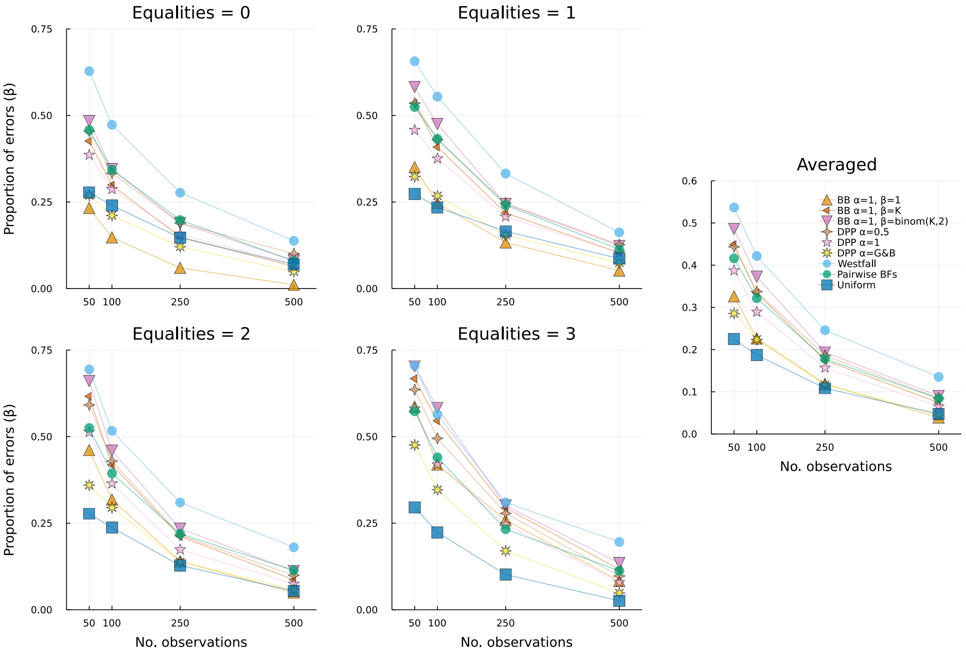


Figure 7.5: Proportion of falsely claiming a difference between two groups when there is none across priors and sample sizes under a model with 0 (top left), 1 (top right), 2 (bottom left), and 3 (bottom right) true inequalities for $K = 5$ groups. The rightmost panel shows the average error rate across inequalities.

Simulation Discussion

Our results show that no single method dominates all others. While the beta-binomial prior with $\beta = 1$ performed best in our initial simulation study described in Section 7.3.1, including models beyond the null and full model showed that this prior performed considerably worse in those settings. The beta-binomial prior with $\beta = K$, $\beta = \binom{K}{2}$, and the DP prior with $\alpha = 0.50$ perform very similarly overall. Importantly, both the method proposed by Westfall et al. (1997) and the pairwise Bayes factors can yield transitivity violations, while explicitly specifying a prior over partitions cannot. For example, we might find that $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$ using pairwise Bayes factors with some threshold, but at the same time conclude that $\mu_1 \neq \mu_3$. This is one key reason why explicitly specifying the prior over partitions is preferable. In the next section, we focus on the beta-binomial prior with $\beta = K$ and apply our method to two examples.

7.4 Applications

In this section, we apply the beta-binomial setup to two examples: testing the (in)equality of proportions and variances, respectively. We have developed a generic Julia package called *EqualitySampler* that utilizes the probabilistic programming framework *Turing* to allow the user to adjust for multiplicity as proposed in this paper.

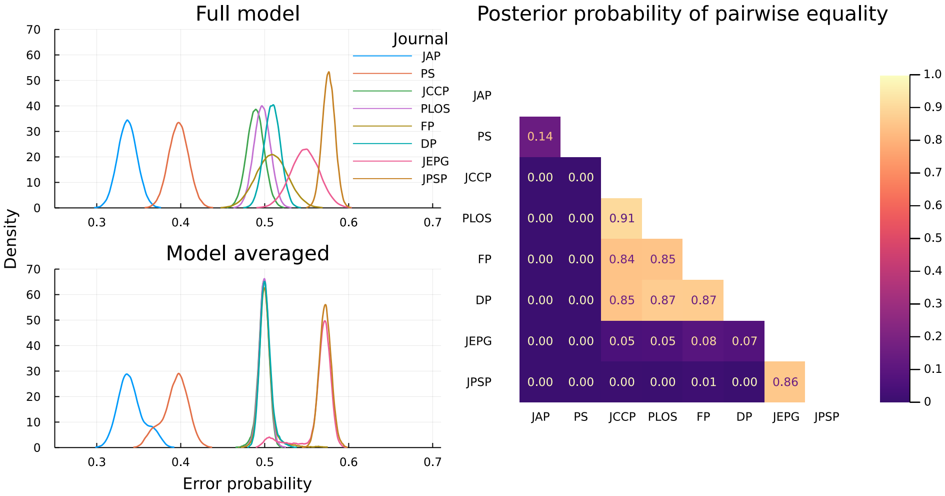
7.4.1 Testing Proportions

Nuijten et al. (2016) investigated a sample of 30,717 articles published between 1985 and 2013 in eight major psychology journals for statistical reporting errors. Our question here is: Which journals make the same amount of errors, and which make more errors? We answer the question using the following model specification. For journal j, denote the number of statistical errors found as e_j and the number of statistical tests analyzed as n_j. We assume that underlying each proportion there is a latent true chance of making an error, θ_j. Thus, we modeled the data as independent binomials, that is, $e_j \sim \text{Binomial}(\theta_j, n_j)$. Next, we specify a hierarchical level over the partitions to assess for which journals the chances of making an error are equal. This leads to the following model specification:

$$\begin{aligned} e_j &\sim \text{Binomial}\left(\theta_j, n_j\right) \\ \theta_j^u &\sim \text{Beta}(1, 1) \\ \theta_j &\leftarrow \text{mean of elements of } \theta_j^u \text{ in the same partition} \\ \rho &\sim \text{beta-binomial}(1, 8) \ . \end{aligned} \tag{7.4.1}$$

The unconstrained chances θ_j^u are assigned beta priors from which — together with the partitions — the possibly constrained chances are created. Two chances θ_i and θ_j are equal if and only if their indices appear in the same partition $\{i, j\} \subseteq \rho_k$ for some k. Note that the model reduces to the full model of independent binomials whenever the partitions state that all elements in $\vec{\theta}$ are distinct. We use a beta-binomial prior with $\alpha = 1$ and $\beta = 8$. The top left panel in Figure 7.6 shows the posterior distributions for the underlying error chance for each journal under a model that assumes that they are all different.

We can see that the posterior distributions for JCCP (green), PLOS (purple), DP (turquoise), and FP (beige) are very close to each other, with FP showing more pronounced uncertainty. The panel below shows the model-averaged posterior distributions, clearly demonstrating a shrinkage effect. The



	JAP	PS	JCCP	PLOS	FP	DP	JEPG	JPSP
JAP								
PS	0.14							
JCCP	0.00	0.00						
PLOS	0.00	0.00	0.91					
FP	0.00	0.00	0.84	0.85				
DP	0.00	0.00	0.85	0.87	0.87			
JEPG	0.00	0.00	0.05	0.05	0.08	0.07		
JPSP	0.00	0.00	0.00	0.00	0.01	0.00	0.86	

Figure 7.6: Left: Posterior means of the full model where all proportions are assumed to be different (top) and posterior means when averaging over all models using a beta-binomial($\alpha = 1$, $\beta = 8$) prior (bottom). Right: Posterior probabilities for pairwise equality across all journals. The abbreviations stand for: *Journal of Applied Psychology* (JAP), *Psychological Science* (PS), *Journal of Consulting and Clinical Psychology* (JCCP), *Public Library of Science* (PLOS), *Developmental Psychology* (DP), *Journal of Experimental Psychology: General* (JEPG), and *Journal of Personality and Social Psychology* (JPSP).

error chances for JAP and PS are pulled toward each other, with JCCP, PLOS, DP, and FP being shrunk towards each other almost completely, similarly to JEPG and JPSP. The right panel in Figure 7.6 gives the posterior distributions for pairwise equality across all journals, reflecting the two main clusters in the model-averaged density plot on the left.

7.4.2 Testing Standard Deviations

Borkenau et al. (2013) studied whether men and women differ in the variability of personality traits. Here we focus on five personality traits (agreeableness, extraversion, openness, conscientiousness, neuroticism) rated by participants' peers in an Estonian sample consisting of $n_1 = 969$ women and $n_2 = 716$ men. Our goal is to assess which personality traits across the sexes can be assumed equal in terms of their variability. This example shows how our methodology can be used to test group differences while taking the multivariate dependency of the outcome measure into account. We build on the parameterization proposed by Dablander et al. (in press), who developed a default Bayes factor test for testing the (in)equality of variances. Let $\vec{y}_1$ and $\vec{y}_2$ denote the five-element vectors of observed data for men and women, respectively, and $K = 10$ be the total number of variables. For each sex $k \in \{1, 2\}$, we have:

$$\begin{aligned}
\vec{Y}_k &\sim \mathcal{N}\left(\vec{\mu}_k, \Sigma_k\right) \\
\vec{\mu}_k &\propto \vec{1} \\
\Sigma_k &= \text{diag}(\vec{\sigma}_k)\, \Omega_k \,\text{diag}(\vec{\sigma}_k) \\
\Omega_k &\sim \text{LKJ}(1) \ \ ,
\end{aligned}$$

where LKJ refers to the Lewandowski-Kurowicka-Joe prior (Lewandowski et al., 2009). To test the equality of variances both between and across groups, we define the ten-variable standard deviation vector $\vec{\sigma} = [\vec{\sigma}_1, \vec{\sigma}_2]$ with $\bar{\sigma}$ denoting the average standard deviation. Following Dablander et al. (in press), we write $\sigma_j = \left(K\vartheta_j\bar{\sigma}\right)^{-1}$, where $\vartheta_j = {}^{\sigma_j}/{\sum_{j=1}^{K}\sigma_j}$ is the relative standard deviation and $\vartheta_K = 1 - \sum_{j=1}^{K-1}\vartheta_j$. To complete the model specification, we write:

$$\begin{aligned}
\sigma_j &= \left(K\vartheta_j\bar{\sigma}_j\right)^{-1} \\
\bar{\sigma}_j &\propto \bar{\sigma}_j^{-1} \\
\vartheta_j &\leftarrow \text{mean of elements of } \vartheta^u \text{ in the same partition} \\
\vec{\vartheta}^u &\sim \text{Dirichlet}(1, \ldots, 1) \\
\rho &\sim \text{beta-binomial}(1, 10) \ \ .
\end{aligned} \tag{7.4.2}$$

Two standard deviations σ_i and σ_j are equal if and only if their indices appear in the same partition $\{i, j\} \subseteq \rho_k$ for some k. When the partition states that

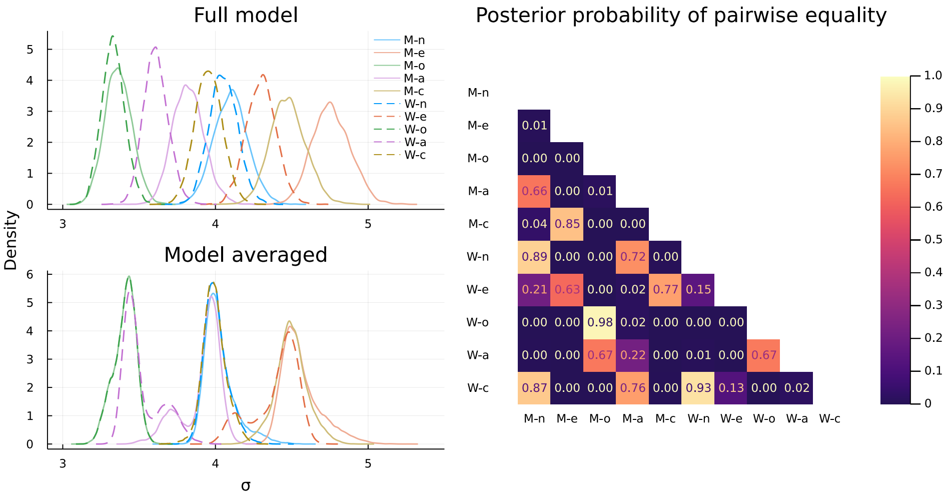


Figure 7.7: Left: Posterior means of the full model where all standard deviations are assumed to be different (top) and posterior means when averaging across all models using a beta-binomial($\alpha = 1$, $\beta = 10$) prior (bottom). Right: Posterior probabilities for pairwise equality across all personality traits. In the abbreviations the first letter stands for *men* (m) or *women* (w). The second letter stands for *neuroticism* (n), *extraversion* (e), *openness* (o), *agreeableness* (a), and *conscientiousness* (c).

all standard deviations are distinct we recover the full model. The top left panel of Figure 7.7 shows the posterior distributions under the full model that assumes all standard deviations are different.

While all posterior distributions lie close to each other, the standard deviations of openness for men and women overlap particularly much. The bottom panel shows the model-averaged posterior distributions, which again demonstrate a shrinkage effect. The right panel of Figure 7.7 shows the posterior probability of pairwise equality across all personality traits for men and women. It appears that there are three clusters: (1) men–openness, women–openness, and women–agreeableness; (2) men–neuroticism, women–neuroticism, women–conscientiousness, and men–agreeableness; (3) men–conscientiousness, men–extraversion, and women–extraversion. However, for the personality traits women–agreeableness, men–agreeableness, and women–extraversion, the evidence is not overwhelming, as indicated by the bimodality in the model-averaged posterior distributions.

7.5 Discussion

Testing the (in)equality between groups while adjusting for multiple comparisons is a core challenge in many applied settings. In this paper, we have proposed a flexible class of beta-binomial priors to penalize multiplicity and make inferences over all possible (in)equalities in relatively general settings. We compared the beta-binomial priors to a Dirichlet process prior suggested by Gopalan and Berry (1998), to a uniform prior, to the method proposed by Westfall et al. (1997), and to an uncorrected method based on pairwise Bayes factors. We also illustrated our method, which is freely available in the Julia package *EqualitySampler*, on two examples.

We found that a beta-binomial prior with $\alpha = 1$ and $\beta \in \{K, \binom{K}{2}\}$ as well as a Dirichlet process prior with $\alpha < 1$ adequately control the familywise error rate, while a uniform prior and using only pairwise Bayes factors, unsurprisingly, do not. We also found that the method proposed by Westfall et al. (1997) compares favorably in terms of error control but not in terms of power. While we have focused on a posterior probability threshold of 0.50 (i.e., a Bayes factor of 1), other thresholds will naturally impact the trade-off between the two types of errors. Importantly, and in contrast to conventional adjustments for multiple comparisons (e.g., Jeffreys, 1961; Westfall et al., 1997), specifying a prior over the partitions allows inferences over all possible (in)equalities. This means that researchers can use the methods we provide to assess not only the probability of pairwise (in)equalities — as is common in standard post-hoc tests for, say, ANOVA — but in fact can make probabilistic statements over any set of (in)equalities they wish to assess. Similarly, the outlined approach also allows for model-averaging, which as we have seen in the applications yields shrinkage of the groups towards each other. Using a prior over partitions further avoids violations of transitivity, i.e. claiming for example that $\mu_1 \neq \mu_3$ while both $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$.

As with any statistical method, there are a number of points to keep in mind. First, while we suggest default values of $\alpha = 1$ and $\beta = K$ for the beta-binomial prior and $\alpha \leq 1$ for the DP prior, researchers may wish to use a more informed prior specification. Values for the prior parameters can be elicited by specifying model priors for two out of the following: the prior on the null model, on the full model, or their ratio. Second, the beta-binomial prior differs from the DP prior in that it assigns models with the same number of partitions the same prior probability, while the DP prior assigns more mass to the model with the larger cluster. It is not obvious which of the two behaviors is more desirable, and it may well depend on the problem under study. Researchers using the methods we have made available should keep this difference in mind, although the extent to which it matters in practice remains to be seen.

There are some practical limitations of our implementation that we leave for future work. We currently do not allow for factorial designs, for example, for which dummy or contrast coding is more natural. The key challenge there is to specify the prior in such a way that it reflects the structure of the experimental design. For the present, we believe that the Bayesian approach outlined in this paper can help applied researchers who wish to compare multiple groups.

Part III

JASP

8

A Tutorial on Bayesian Multi-Model Linear Regression with BAS and JASP

Linear regression analyses commonly involve two consecutive stages of statistical inquiry. In the first stage, a single 'best' model is defined by a specific selection of relevant predictors; in the second stage, the regression coefficients of the winning model are used for prediction and for inference concerning the importance of the predictors. However, such second-stage inference ignores the model uncertainty from the first stage, resulting in overconfident parameter estimates that generalize poorly. These drawbacks can be overcome by model averaging, a technique that retains all models for inference, weighting each model's contribution by its posterior probability. Although conceptually straightforward, model averaging is rarely used in applied research, possibly due to the lack of easily accessible software. To bridge the gap between theory and practice, we provide a tutorial on linear regression using Bayesian model averaging in `JASP`, based on the BAS package in `R`. Firstly, we provide theoretical background on linear regression, Bayesian inference, and Bayesian model averaging. Secondly, we demonstrate the method on an example data set from the World Happiness Report. Lastly, we discuss limitations of model averaging and directions for dealing with violations of model assumptions.

This chapter is published as: van den Bergh, D., Clyde, M. A., Komarlu Narendra Gupta, A. R., de Jong, T., Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2021a). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, 1–21.

Linear regression is a standard statistical procedure in which one continuous variable (known as the dependent, outcome, or criterion variable) is being accounted for by a set of continuous predictor variables (also known as independent variables, covariates, or predictors). For concreteness, consider a researcher who is interested in predicting people's happiness using a number of country-specific demographic indicators such as Gross Domestic Product (GDP), public safety, life expectancy, and many others. When all available predictors are included in the regression equation, the resulting model will generally overfit the data, the estimates of the regression coefficients will be unreliable, and the results will generalize poorly to other data sets (e.g., Myung, 2000). Therefore, most regression analyses start by reducing the set of initial predictors to a relevant subset. The challenge of identifying a good subset is known as the model selection or variable selection problem. For instance, a variable selection procedure may suggest that only wealth and life expectancy are needed to predict happiness. Once the relevant subset has been identified, the associated regression model can be used to assess the magnitude of the relations between the criterion variable and the selected subset of predictors (e.g., how much we expect happiness to change per unit of change in wealth).

Although common practice, the two-step procedure has been known to be problematic for over 25 years (e.g., Hurvich & Tsai, 1990; Miller, 1990). Specifically, the second step in the two-step procedure ignores the uncertainty associated with the first step, that is, the uncertainty with which the model of interest (i.e., the subset of predictors) was obtained. Consequently, inference from two-step methods has been shown to be misleading (Draper, 1995) and result in overconfident parameter estimates and biased inference (Burnham & Anderson, 2002, Ch. 1.7). As summarized by Claeskens and Hjort (2008, Ch 7.4, p. 199):

> "*'Standard practice' has apparently become to use a model selection technique to find a model, after which this part of the analysis is conveniently forgotten, and inference is carried out as if the selected model had been given a priori. This leads to too optimistic tests and confidence intervals, and generally to biased inference statements.*" (italics in original)

The principled alternative to the two-step procedure is multi-model inference. Instead of settling, perhaps prematurely, on a single model for inference, multi-model inference retains all models and calculates for each model a weight that indicates the degree to which the data support that model. These weights are usually a function of the posterior model probabilities, which represent the relative probability in favor of each model after the data are observed (Hoeting et al., 1999; Raftery et al., 1997). At the same time that the model weights are being obtained, parameter estimates are calculated for each model. Then, instead of basing all of our inferences on a single model, we can take into account all of the models simultaneously. For example, in order to predict a set of new observations we first generate predictions from the individual models and then average these predictions using the posterior model probabilities as weights. This ensures our final prediction for new observations reflects our uncertainty across the entire model space (Claeskens & Hjort, 2008, Ch. 7). In other words, multi-model inference accomplishes variable selection and parameter estimation simultaneously instead of sequentially.

Despite the advantages of multi-model inference (e.g., Burnham et al., 2011; Hinne et al., 2020; Hoeting et al., 1999) and its successes in fields such as machine learning (Breiman, 2001), cosmology (Trotta, 2008), and climate prediction (Tebaldi & Knutti, 2007), the procedure has been applied only rarely in psychology (but see e.g., Gronau, Van Erp, et al., 2017; Kaplan & Lee, 2016). The lack of multi-model inference in psychological science may be due in part to the perceived lack of user-friendly software that executes the analysis, as well as a dearth of tutorial-style explanations that allow psychologists to interpret the results of multi-model inference.

This aim of this paper is to bridge the gap between theory and practice by providing a tutorial on Bayesian multi-model inference, with an emphasis on user-friendly software to execute the analysis. First, we briefly provide theoretical background on linear regression, Bayesian inference, and Bayesian multi-model inference. Next we demonstrate the method in action using the BAS `R` package (Clyde, 2018) as implemented in `JASP` (JASP Team, 2022), an open source software program with a graphical user interface. The paper concludes with a summary and a discussion about pitfalls of regression modeling.

8.1 Theoretical Background

Before demonstrating Bayesian multi-model linear regression for a concrete data set we first introduce some basic theory. The impatient reader may skip this section. Below we first introduce linear regression, its assumptions, and the most common measure of effect size, R^2. We then briefly describe Bayesian inference and finally introduce multi-model inference.

8.1.1 Linear Regression

The most common definition of multiple regression is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{8.1.1}$$

where i refers to the scores of the i^{th} subject and p to the total number of predictors. The intercept is represented by β_0, and the linear effects between criterion and predictor variables are given by the regression coefficients $\beta_1, \ldots,$ β_p. The residuals (ϵ_i) are assumed to be normally distributed with mean 0 and unknown variance σ^2. The predictors ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_p$) are usually centered (i.e., modeled with their mean subtracted, for example $\beta_1 \left(x_{i1} - \overline{x}_1\right)$) so that inference about the intercept is independent of which predictors are included in the model. We will refer to collections of parameters or data points (vectors) using bold notation (e.g., $\boldsymbol{y}$ denotes y_1, y_2, …, y_n).

From the definition of linear regression, it is evident that the model space can be enormous; consequently, linear regression presents a multi-model problem. With p predictors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$, each of which can be included or excluded from the model, the total model space consists of 2^p members (e.g., with 10 predictors, there are 1024 different models to consider; with 15 predictors, the space grows to $32,768$ models). If interaction effects are considered, the model space grows even more rapidly.

Results from a linear regression analysis can be misleading if its assumptions are violated. The key assumption of linear regression is that the residuals are normally distributed. Introductory texts often mention other assumptions, but these assumptions generally concern specific violations of normality. We recommend three visual checks for assessing normality. As the name linear regression suggests, the relation between the predictor variables and the criterion variable should be approximately linear. Therefore, the first visual check we recommend is examining a scatter plot of the criterion and predictor variables. For example, suppose we wish to predict Happiness using Wealth. We might observe that the distribution of Wealth is right skewed and that the relation between Happiness and Wealth is non-linear. Such deviations from linearity can be corrected using, for instance, a log-transformation. Note that because of such transformations, linear regression analyses can detect more than just linear trends. The relation between Happiness and Wealth is shown in Figure 8.1.

Second, we recommend examining a Q-Q plot to assess the normality of the residuals. A Q-Q plot shows the quantiles of a theoretical normal distribution against the observed quantiles of the residuals. If the observed residuals are approximately normal, then all points in the plot fall approximately on a straight line. However, not all deviations from normality are easy to detect in a Q-Q plot. For instance, a Q-Q plot does not clearly show if the residuals are heteroscedastic, that is, the variance of the residuals is not constant across predictions. Therefore, our third recommendation is to plot a model's predictions against a model's residuals, which is a common visualization to assess heteroscedasticity and nonlinearity. To illustrate, we again predict Happiness with Wealth as measured in GPD. The left panel of Figure 8.2 shows a Q-Q

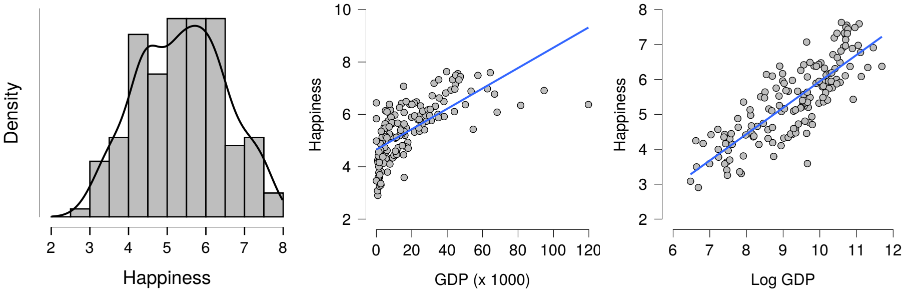


Figure 8.1: Example of a non-linear relationship between Happiness and Wealth, measured in terms of GDP. The left panel shows the density estimate for Happiness, the middle and right panel relate Happiness (y-axis) to GDP and log-transformed GDP (x-axes), respectively.

plot of theoretical against observed residuals and indicates little deviation from normality. However, the right panel of Figure 8.2 visualizes the model's predictions against the model's residuals and suggests that the variance of the prediction error depends on the model's predictions. For example, the residuals for a prediction of 5 are much more spread out than the residuals for a prediction of 6. In the right panel, the red line is a smoothed estimate of the mean at each point, obtained with local polynomial regression (Cleveland et al., 1992). If the red line were horizontal with intercept zero, this would indicate that there is no structure left in the residuals that could be captured by the model (e.g., with interaction effects or higher-order polynomial terms). However, here the red line varies as a function of the predictions, most likely because the relation between predictor and criterion is non-linear. Furthermore, the variance of the residuals differs across the predictions. This indicates that the residuals are heteroscedastic. A linear regression of Happiness predicted by log-transformed GDP yields residuals that are better in agreement with the assumptions of linear regression (see Appendix, Figure G.1).

After applying the regression model of interest and having confirmed that the assumptions are not badly violated, it is recommended to assess model fit. Model fit indices provide an idea about how well the model describes the data. Among the many model fit indices, the most common is the coefficient of determination R^2 (Olive, 2017, p. 31), defined as

$$R^2_{\mathcal{M}_j} = \text{Cor}\left(\boldsymbol{y}, \, \hat{\boldsymbol{y}} \mid \mathcal{M}_j\right)^2 . \tag{8.1.2}$$

$R^2_{\mathcal{M}_j}$ is the proportion of variance of the criterion variable $\boldsymbol{y}$ that is explained by model $\mathcal{M}_j$. The explained variance is computed by squaring the sample correlation between the observations $\boldsymbol{y}$ and the predictions $\hat{\boldsymbol{y}}$ of $\mathcal{M}_j$. Usually, the term $\mathcal{M}_j$ is omitted for brevity. Since R^2 is the square of a correlation it

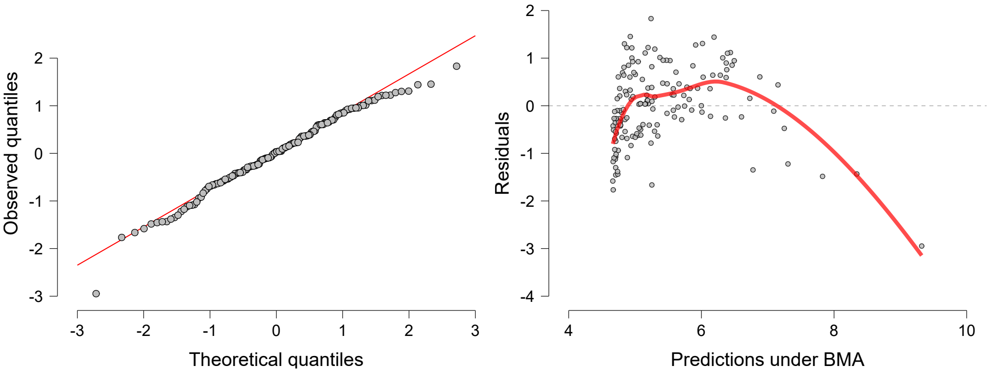


Figure 8.2: Assumptions checks for a linear regression where Happiness is predicted from Wealth, measured in terms of GDP. The left panel shows a Q-Q plot of the theoretical quantiles expected under a normal distribution (x-axis) against the quantiles of the observed residuals obtained from Bayesian Model Averaging (BMA; y-axis). The residuals appear approximately normally distributed. The right panel plots the predictions under BMA (x-axis) against the residuals (y-axis). Figures from JASP.

always lies between 0 (poor model fit) and 1 (perfect model fit). It should be stressed that R^2 is *not* a good measure for model comparison because it does not penalize models for complexity: when additional predictors are added to a model, R^2 can only increase. Therefore, R^2 will always favor the most complex model. However, the most complex model often fits the data too well, in the sense that idiosyncratic noise is misperceived to be systematic structure. In other words, complex models are prone to overfit the data (e.g., Hastie et al., 2008, Ch. 7; Myung and Pitt, 1997; Vandekerckhove et al., 2015). Because models that overfit the data treat irreproducible noise as if it were reproducible signal, predictive performance for new data suffers. Altogether, this makes R^2 unsuitable for model selection, unless the competing models have the same number of predictors.

8.1.2 Bayesian Inference

The next sections provide a brief introduction to Bayesian statistics. For accessible, in-depth tutorials and an overview of the literature we recommend the recent special issue in *Psychonomic Bulletin & Review* (Vandekerckhove et al., 2018).

Bayesian Parameter Estimation

Given a specific model $\mathcal{M}_j$ –in regression, a particular subset of predictors– we start a Bayesian analysis by defining prior beliefs about possible values for the parameters (e.g., the regression coefficients). This belief is represented as a probability distribution; ranges of likely values have more prior probability and ranges of less likely values have less prior probability.

As soon as data $\mathcal{D}$ are observed, Bayes' theorem (Equation 8.1.3) can be used to update the prior distribution to a posterior distribution:

$$\underbrace{p(\boldsymbol{\beta} \mid \mathcal{D}, \mathcal{M}_j)}_{\text{Posterior}} = \overbrace{p(\boldsymbol{\beta} \mid \mathcal{M}_j)}^{\text{Prior}} \times \frac{\overbrace{p(\mathcal{D} \mid \boldsymbol{\beta}, \mathcal{M}_j)}^{\text{Likelihood}}}{\underbrace{p(\mathcal{D} \mid \mathcal{M}_j)}_{\substack{\text{Marginal}\\ \text{Likelihood}}}}. \tag{8.1.3}$$

Equation 8.1.3 shows that our prior beliefs are adjusted to posterior beliefs through an updating factor that involves the likelihood (i.e., predictive performance for specific values for β) and the marginal likelihood (i.e., predictive performance across all values for β): values for β that predicted the data better than average receive a boost in plausibility, whereas values of β that predicted the data worse than average suffer a decline (e.g., Wagenmakers et al., 2016). Equation 8.1.3 also shows that the posterior distribution is a compromise between the prior distribution (i.e, our background knowledge) and the data (i.e., the updating factor). The updating process is visualized in Figure 8.3. Note that the impact of the prior on the posterior becomes less pronounced when sample size increases. In large samples, the posterior is often dominated by the likelihood and the posterior is practically independent of the prior (Wrinch & Jeffreys, 1919). In addition, with more data the posterior distribution becomes increasingly peaked, reflecting the increased certainty about the value of the parameters.

Bayesian Model Selection

The parameter estimation procedure provides us with posterior distributions for parameter values conditional on a given model $\mathcal{M}_j$. When multiple models are in play, we can extend Bayes' theorem and use the data to update the relative plausibility of each of the candidate models. For the case of two models, $\mathcal{M}_0$ and $\mathcal{M}_1$, Equation 8.1.4 shows how the prior model odds (i.e., the relative plausibility of $\mathcal{M}_0$ and $\mathcal{M}_1$ before seeing the data) are updated to posterior model odds (i.e., the relative plausibility of $\mathcal{M}_0$ and $\mathcal{M}_1$ after seeing the data). The change from prior to posterior odds is given by the *Bayes factor* (e.g., Jeffreys, 1961; Kass & Raftery, 1995), which indicates the

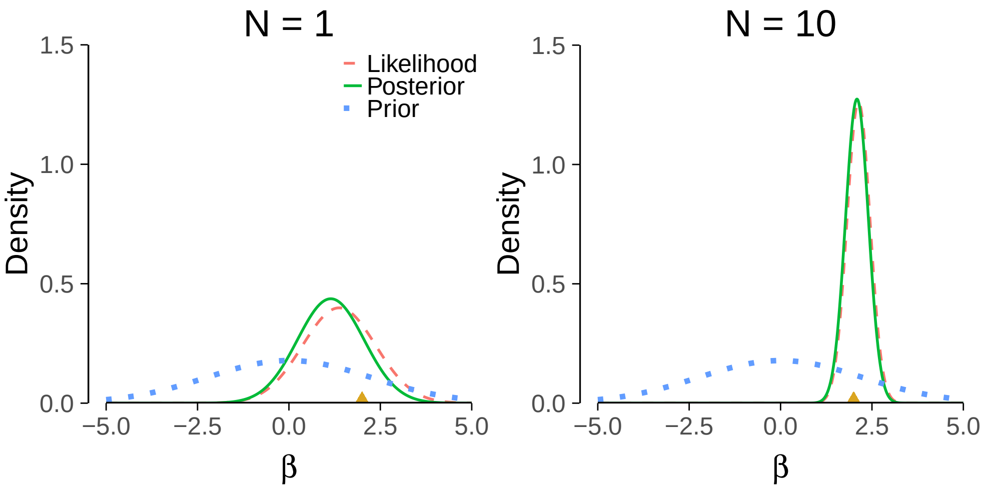


Figure 8.3: Illustration of Bayesian updating using Bayes' theorem for a single observation (left panel) and ten observations (right panel). The 'true' value is 2 and is indicated by the gold triangle on the x-axes. Note that (1) the posterior depends less on the prior as more data are observed; (2) the variance (width) of the posterior decreases with sample size. In other words, we become more certain of our estimates as we observe more data. In the right panel, the likelihood was normalized for illustrative purposes. This example is based on normally distributed data with unknown mean and known variance (for derivations, see Murphy, 2007).

models' relative predictive performance for the data at hand (i.e., the ratio of marginal likelihoods):

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \mathcal{D})}{p(\mathcal{M}_0 \mid \mathcal{D})}}_{\text{Posterior model odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}_{\text{Prior model odds}} \times \underbrace{\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_0)}}_{\substack{\text{Bayes factor} \\ \text{BF}_{10}}}. \tag{8.1.4}$$

When the Bayes factor BF_{10} is 4 this indicates that the data are 4 times more likely under $\mathcal{M}_1$ than $\mathcal{M}_0$. The Bayes factor subscripts indicate which model is in the numerator and denominator; for instance, if $\text{BF}_{10} = 0.20$, then 1 / $\text{BF}_{10} = \text{BF}_{01} = 5$, which means that the data are 5 times more likely under $\mathcal{M}_0$ than under $\mathcal{M}_1$ (Jeffreys, 1939). There exist several categorization schemes to quantify the evidence associated with particular ranges of values (e.g., Jeffreys, 1961; Kass & Raftery, 1995). Table 8.1 provides one such scheme.

With more than two candidate models in the set, the posterior model probability for model $\mathcal{M}_j$ is given by

Table 8.1: A scheme for categorizing the strength of a Bayes factor (from Lee and Wagenmakers, 2013, based on Jeffreys, 1961). Note that the Bayes factor is a continuous measure of evidence and that the thresholds provided here (and in other schemes) are only meant as a heuristic guide to facilitate interpretation and not as a definite cutoff.

Bayes factor BF_{10}	Interpretation
> 100	Extreme evidence for $\mathcal{M}_1$
$30 - 100$	Very strong evidence for $\mathcal{M}_1$
$10 - 30$	Strong evidence for $\mathcal{M}_1$
$3 - 10$	Moderate evidence for $\mathcal{M}_1$
$1 - 3$	Anecdotal evidence for $\mathcal{M}_1$
1	No evidence
$^1/_3 - 1$	Anecdotal evidence for $\mathcal{M}_0$
$^1/_{10} - {}^1/_3$	Moderate evidence for $\mathcal{M}_0$
$^1/_{30} - {}^1/_{10}$	Strong evidence for $\mathcal{M}_0$
$^1/_{100} - {}^1/_{10}$	Very strong evidence for $\mathcal{M}_0$
$< {}^1/_{100}$	Extreme evidence for $\mathcal{M}_0$

$$p(\mathcal{M}_j \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i)p(\mathcal{M}_i)}.$$

This can also be written as a function of the Bayes factor relative to the null model:

$$p(\mathcal{M}_j \mid \mathcal{D}) = \frac{\text{BF}_{j0}\;\, p(\mathcal{M}_j)}{\sum_i \text{BF}_{i0}\;\, p(\mathcal{M}_i)}.$$

The change from prior to posterior model odds quantifies the evidence $\text{BF}_{\mathcal{M}_j}$ that the data provide for a particular model j. The prior model odds are given by $^{p(\mathcal{M}_j)}/_{1-p(\mathcal{M}_j)}$ and the posterior model odds are given by $^{p(\mathcal{M}_j \mid \mathcal{D})}/_{1-p(\mathcal{M}_j \mid \mathcal{D})}$. The change in odds is obtained by dividing the posterior model odds by the prior model odds:

$$\text{BF}_{\mathcal{M}_j} = \frac{p(\mathcal{M}_j \mid \mathcal{D})}{1 - p(\mathcal{M}_j \mid \mathcal{D})} \Bigg/ \frac{p(\mathcal{M}_j)}{1 - p(\mathcal{M}_j)}.$$

Bayes factors generally depend on the prior distribution for the parameter values. In contrast to estimation, the data do not overwhelm the prior because the Bayes factor quantifies relative predictive performance of two models on a data set.[1] This is desirable because complex models usually yield many poor predictions and therefore the Bayes factor inherently penalizes complexity and favors parsimony (Jeffreys, 1961). However, without reliable information suitable for constructing a prior, the relation between Bayes factors and priors introduces the need for default prior distributions.

There are two types of prior distributions that need to be decided upon. The first type of prior distribution is the *model prior*, which assigns a prior probability to each model that is considered. For the time being, we only consider a uniform model prior so that all models are a-priori equally likely. Alternative model priors are discussed in the section *Prior Sensitivity*.

The second type of prior distribution is the prior on parameters. A popular choice of default prior distributions for parameters β in linear regression is the Jeffreys–Zellner–Siow (JZS) prior (i.e., a multivariate Cauchy distribution on the beta coefficients) which is also used in the implementation shown later. The JZS prior fulfills several desiderata (see Liang et al., 2008; Zellner, 1986; Zellner and Siow, 1980 for information on the JZS-prior, see Rouder and Morey, 2012 for default priors in Bayesian linear regression, and see Ly et al., 2016a for a general introduction on default Bayes factor hypothesis tests). An example of such a desideratum is that the Bayes factor is the same regardless of the units of measurement (e.g., the Bayes factor is the same when response time is measured in milliseconds or years; for more information see Bayarri et al., 2012). This desideratum is satisfied by assigning a Jeffreys prior to the residual variance σ^2, that is, $p(\sigma^2)$ is proportional to $^1/_{\sigma^2}$.

Other methods included in JASP are the Akaike Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), the g-prior (Zellner, 1986), the hyper-g prior (Liang et al., 2008), the hyper-g-Laplace prior which is the same as the hyper-g prior but uses a Laplace approximation, and the hyper-g-n prior which uses a hyper-$^g/_n$ prior (Liang et al., 2008). In addition, two methods are available that use a g-prior and automatically choose a value for g. Empirical Bayes "global" uses an EM algorithm to find a suitable value for g while empirical Bayes "local" uses the maximum likelihood estimate for each individual model as value for g (Clyde & George, 2000). We revisit the possible use of these alternative methods when we discuss robustness.

8.1.3 Bayesian Multi-Model Inference

As before, assume that there are multiple models in play, each with their own set of predictors. In the previous section we have seen that the posterior model probabilities can be obtained by assessing each model's plausibility and

[1] As the words imply, predictions follow from the prior distribution; postdictions follow from the posterior distribution.

predictive performance, relative to that of the other models in the set. When the results point to a single dominant model, then it is legitimate to consider only that model for inference. When this is not the case, however, inference about the predictors needs to take into account multiple models at the same time. We consider two important questions: (1) what predictors should be included to account for the dependent variable? and (2) what have we learned about the regression coefficients for the predictors? In multi-model inference, these questions can be addressed by summing and averaging across the model space, respectively.

First, consider the question 'if we want to predict Happiness, do we need the predictor Wealth?' There may be thousands of regression models, half of which include Wealth as a predictor, and half of which do not. In BMA we can quantify the overall support for the predictor Wealth by summing all posterior model probabilities for the models that include Wealth:

$$p(\text{incl}_{\beta_j} \mid \mathcal{D}) = \sum_{\mathcal{M}_j : \beta_j \in \mathcal{M}_j} p(\mathcal{M}_j \mid \mathcal{D})$$

If the summed prior probability of models including Wealth is 0.50, and the summed posterior probability is 0.95, then the inclusion Bayes factor is 19. That is:

$$\frac{p(\text{incl}_{\beta_j} \mid \mathcal{D})}{p(\text{excl}_{\beta_j} \mid \mathcal{D})} = \frac{p(\mathcal{D} \mid \text{incl}_{\beta_j})}{p(\mathcal{D} \mid \text{excl}_{\beta_j})} \frac{p(\text{incl}_{\beta_j})}{p(\text{excl}_{\beta_j})}$$

Second, consider the question 'what have we learned about the regression coefficient for the predictor Wealth?' In the models that do not feature Wealth, this coefficient can be considered zero; in the models that do feature Wealth, the coefficient has a posterior distribution, but a different one for each model. In BMA, we can provide an overall impression of our knowledge about the coefficient by averaging the parameter values across all of the models, using the posterior model probabilities as weights (e.g., Ghosh, 2015; Raftery et al., 1997). Intuitively, one can first sample a model (using the posterior model probabilities) and then, from that model, draw a value of the regression coefficient from the posterior distribution for that model; repeating this very many times gives a model-averaged posterior distribution for the regression coefficient of interest. Specifically, we have:

$$p(\beta \mid \mathcal{D}) = \sum_j p(\beta \mid \mathcal{D}, \mathcal{M}_j)\, p(\mathcal{M}_j \mid \mathcal{D})$$

The same procedure for sampling from the posterior distribution of the regression coefficients can be used to obtain a distribution over model-based predictions. Letting $\hat{y}_i$ denote a prediction for outcome i we obtain:

$$p(\hat{y}_i \mid \mathcal{D}) = \sum_j p(\hat{y}_i \mid \mathcal{D}, \mathcal{M}_j)\, p(\mathcal{M}_j \mid \mathcal{D})$$

Here, one may use the observed values for the predictors to obtain fits for the observed values of the criterion variable, or one can use new values for the predictors to obtain predictions for unseen values of the criterion variable. Note that the predictions and the residuals are random variables endowed with probability distributions, rather than single values.

A complementary method is to base all inference on the *median probability model* (Barbieri, Berger, et al., 2004) which includes all predictors that have posterior inclusion probabilities larger than or equal to 0.5. This method is implemented both in BAS and in JASP.

Although BMA is theoretically straightforward, considerable practical challenges need to be overcome. The main challenge is that the model space can be truly enormous, and consequently even advanced computational methods can grind to a halt. Fortunately, the computational challenge surrounding Bayesian multi-model inference in linear regression has been mostly overcome by a recent method called Bayesian Adaptive Sampling (BAS Clyde et al., 2011b). In principle, BAS tries to enumerate the model space if $p \leq 20$. However, if the model space is too large to enumerate –when $p > 20$ implying that there are more than $1,048,576$ models to consider– BAS uses an efficient method for sampling from the model space without replacement. An open-source implementation of BAS is available for R (R Core Team, 2022; package 'BAS', Clyde, 2018) and the methodology is also accessible with a graphical user interface in JASP JASP Team, 2022.

8.2 Example: World Happiness Data

To showcase Bayesian multi-model inference for linear regression we consider data from the World Happiness Report of 2018. The data set can be obtained from the appendix of http://worldhappiness.report/ed/2018/. An annotated `.jasp` file of the analysis detailed below can be found at https://osf.io/5dmj7/. The goal of the analysis is to examine which variables are related to Happiness, and what is the strength of the relation. First we briefly describe the data set.

The World Happiness Data is put together yearly by Gallup, a research-based consulting company. Gallup regularly conducts public opinion polls and annually conducts interviews with a large number of inhabitants of many different countries.[2] The happiness of the interviewees was assessed with the Cantril Self-Anchoring Striving Scale (Glatzer & Gulyas, 2014). In addition, interviewees were asked about a variety of topics and the obtained data are distilled into six variables that may relate to happiness. A description of these six variables is given in Table 8.2.

Table 8.2: Description of the predictor variables for the Gallup World Happiness Data. For a more detailed description of the variables see technical box 1 of Gallop's complete report.

Predictor	Abbreviation	Description
GDP per Capita	W	The relative purchasing power of inhabitants of a country, based on data from the World Bank.
Life Expectancy	Le	Life expectancy based on data from the World Health Organization.
Social Support	Ss	The nation-wide average of responses to the question: 'If you were in trouble, do you have relatives or friends you can count on to help whenever you need them, or not?'
Freedom	F	The nation-wide average to the question: 'Are you satisfied or dissatisfied with your freedom to choose what you do with your life?'
Generosity	Ge	The nation-wide average 'Have you donated to a charity in the last month?'
Perception of Corruption	Poc	The nation-wide average to the questions 'Is corruption widespread throughout the government or not?' and 'Is corruption widespread within businesses or not?'.

We first analyze the data using a standard Bayesian multi-model approach, which is then extended to deal with interaction effects, nuisance variables included in all models, and robustness checks.

Before carrying out any analyses it is critical to check the model assumptions. We investigate the assumption of linearity by plotting the entire set of independent variables against the dependent variable, as shown in Figure 8.4. To replicate Figure 8.4, open `JASP` and load the data, go to `Descriptives`,

[2] Specific information about the data collection can be found on their website http://www.gallup.com/178667/gallup-world-poll-work.aspx. Gallop's complete report can be downloaded from http://worldhappiness.report/ed/2018/.

first drag your dependent variable and then all independent variables.[3] Then under `Plots` click `Correlation plot`.

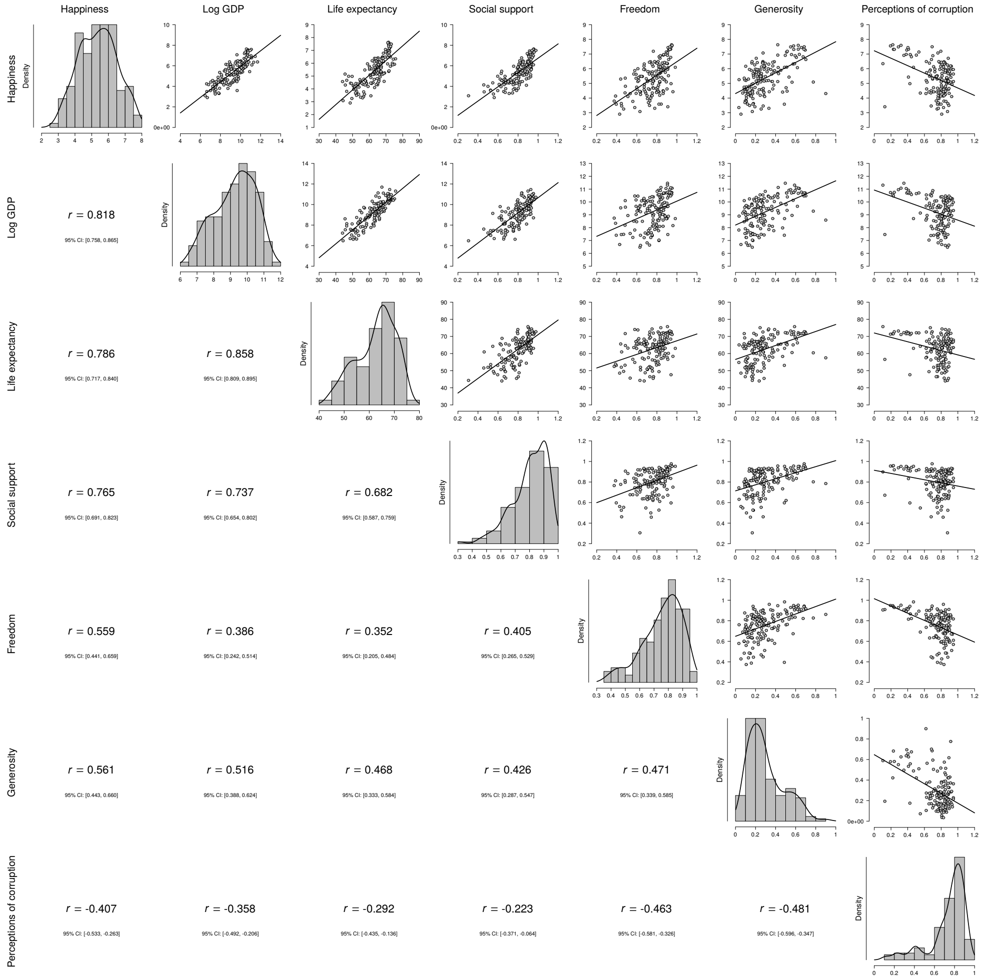
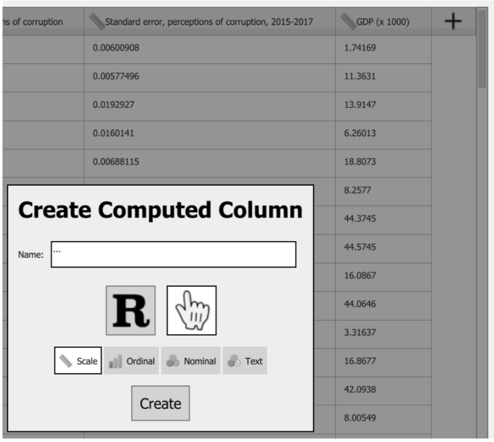


Figure 8.4: A matrix-plot of all variables in the World Happiness Data. The diagonal plots are the density estimates of the individual variables. The above-diagonal plots are pairwise scatter plots of two variables, where the straight line represent the correlation between them. In the first row, Happiness score (y-axes) is plotted against all independent variables (x-axes). Below the diagonal the Pearson correlations are displayed. All relations appear approximately linear by eye. Figure from JASP.

Figure 8.4 shows that all relations between the covariates and Happiness are approximately linear. Initially, the relation between Happiness and Wealth

[3] All `JASP` commands in the input menu are typeset `like this.`

was nonlinear (see Figure 8.1), but after log-transforming Wealth this assumption no longer appears violated (as shown in Figure 8.4). Transforming a variable in `JASP` can be done by going to the data view, scrolling all the way to the right and selecting `Compute Columns`. Next, we can create a new variable, either using a drag and drop scheme or using R-code. This is shown in Figure 8.5.



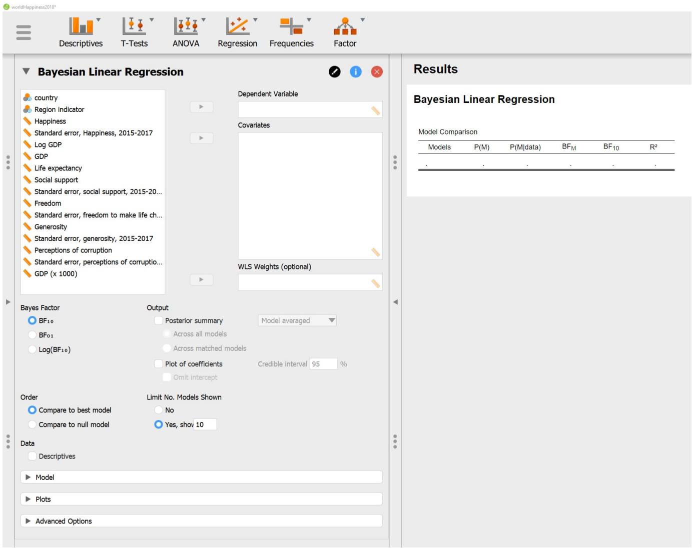
ns of corruption	Standard error, perceptions of corruption, 2015-2017	GDP (x 1000)
	0.00600908	1.74169
	0.00577496	11.3631
	0.0192927	13.9147
	0.0160141	6.26013
	0.00688115	18.8073
		8.2577
		44.3745
		44.5745
		16.0867
		44.0646
		3.31637
		16.8677
		42.0938
		8.00549

Figure 8.5: Compute a new column in JASP by clicking on the '+' in the top right of the data view.

The other key assumption –normally distributed residuals– can only be studied after executing the analysis. To execute the analysis in `JASP`, we go to the Regression menu and click on `Bayesian Linear Regression`. Figure 8.6 shows the resulting interface. We enter the data by dragging `Happiness` to the box labeled `Dependent Variable` and by dragging the independent variables to the box labeled `Covariates`. As soon as the data are entered the analysis is carried out and the table on the right of Figure 8.6 is filled out. Before interpreting the results we assess whether the residuals are approximately normally distributed. To do so, we go to `Plots` and check `Residuals vs. fitted`. This produces the left panel of Figure 8.7, which shows there is still structure in the residuals that is not captured by the model. We included a two-way interactions between Life expectancy and Social support.[4] This is motivated by the following comment in Gallop's report (page 21):

[4] The model space considered should be predetermined and preferably preregistered before commencing with the analysis. We enlarge the model space here to meet the model assumptions. Strictly speaking, the results should be viewed as exploratory.

> "*There are also likely to be vicious or virtuous circles, with two-way linkages among the variables. For example, there is much evidence that those who have happier lives are likely to live longer, be more trusting, be more cooperative, and be generally better able to meet life's demands. This will feed back to improve health, GDP, generosity, corruption, and sense of freedom.*" (original in italics)



▼ Bayesian Linear Regression

country
Region indicator
Happiness
Standard error, Happiness, 2015-2017
Log GDP
GDP
Life expectancy
Social support
Standard error, social support, 2015-20...
Freedom
Standard error, freedom to make life ch...
Generosity
Standard error, generosity, 2015-2017
Perceptions of corruption
Standard error, perceptions of corruptio...
GDP (x 1000)

Dependent Variable

Covariates

WLS Weights (optional)

Bayes Factor
● BF₁₀
○ BF₀₁
○ Log(BF₁₀)

Output
□ Posterior summary   Model averaged
○ Across all models
○ Across matched models
□ Plot of coefficients   Credible interval 95 %
□ Omit intercept

Order
● Compare to best model
○ Compare to null model

Limit No. Models Shown
○ No
● Yes, show 10

Data
□ Descriptives

Results

Bayesian Linear Regression

Model Comparison

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
.	.	.	.	.	.

Figure 8.6: Screenshot of Bayesian linear regression in `JASP`. The left panel shows the input fields; once these are populated, output will appear in the panel on the right.

After confirming that the assumptions of linear regression have been met, we can investigate the results. No further action is required; as soon as the data were entered, `JASP` executed the analysis and displayed the results in an output table. The results for the ten models with the highest posterior probability are shown in Table 8.3.

Table 8.3 shows that the ten best models all contain Life expectancy, Social support, and Freedom, which suggests that these predictors are important to account for Happiness. Also, note that the Bayes factor BF_{01}, which quantifies a model's relative predictive performance, does not always prefer models with higher explained variance R^2, which quantifies a model's goodness-of-fit. For instance, R^2 is necessarily highest for the full model that contains all seven predictors (row 5 in Table 8.3); however, the Bayes factor indicates that the predictive performance of this relatively complex model is about 66 times worse than that of the model that contains only Wealth, Life Expectancy,

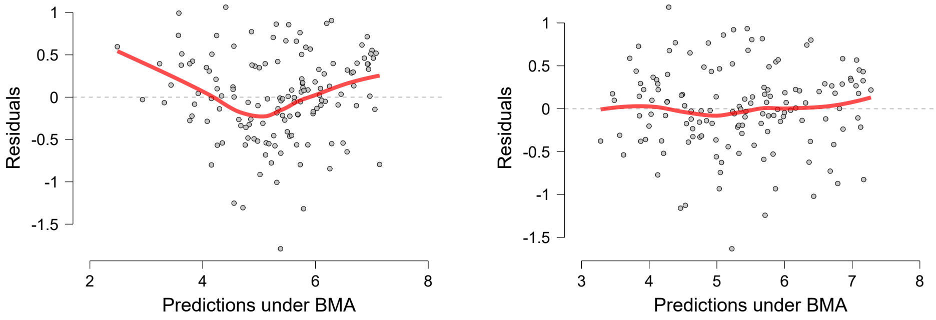


Figure 8.7: Residuals vs Predictions for the World Happiness data set for the model without (left panel) and with (right panel) the interaction effect of Life expectancy and Social support. The red line is a smoothed estimate of the mean at each point and is ideally completely flat. Figures from JASP.

Social support, Freedom, and the interaction between Life expectancy and Social support.

With many different models it can be challenging to quantify the relevance of individual predictors by showing all models as in Table 8.3 (and its complete version with all 80 models). In model-averaging, the solution is to take into account all models simultaneously. This can be accomplished in `JASP` by ticking `Posterior summary` in the input panel and selecting the option `Model averaged`. The output, shown here in Table 8.4, provides a summary of the predictor inclusion probabilities and the posterior distributions averaged across all models.

Table 8.4 confirms our initial impression about the importance of Wealth, Life expectancy, Social Support, Freedom, and the interaction between Life expectancy and Social Support. Each of these predictors are relevant for predicting Happiness, as indicated by the fact that the posterior inclusion probabilities (0.962, 1.000, 1.000, 1.000, and 0.998 respectively) are all near 1.[5] On the other hand, there is evidence against the relevance of Generosity and Perception of Corruption: the data lowered the inclusion probabilities from 0.5 to about 0.1. The median probability model (i.e., the model that includes all predictors with a posterior inclusion probability larger than 0.5, Barbieri, Berger, et al., 2004) consists of Wealth, Life expectancy, Social support, Freedom, and the interaction between Life expectancy and Social support. To obtain the posterior summary for the median probability model, click on the menu that says `Model averaged` and change it to `Median model`.

[5] Although JASP rounds the posterior inclusion probabilities to 1, they never equal 1 exactly.

Table 8.3: The 10 best models from the Bayesian linear regression for the Gallup World Happiness Data. The leftmost column shows the model specification, where each variable is abbreviated as in Table 8.2. The second column gives the prior model probabilities; the third the posterior model probabilities; the fourth the change from prior to posterior model odds; the fifth the Bayes factor of the best model over the model in that row; and the last the R^2, the explained variance of each model. Results for all 80 models are presented in the appendix, Table G.1.

Models	$P(\mathcal{M})$	$P(\mathcal{M} \mid \mathcal{D})$	$\text{BF}_{\mathcal{M}}$	BF_{01}	R^2
W + Le + Ss + F + Le * Ss	0.013	0.759	248.244	1.000	0.821
W + Le + Ss + F + Ge + Le * Ss	0.013	0.097	8.531	7.783	0.822
W + Le + Ss + F + Poc + Le * Ss	0.013	0.093	8.101	8.157	0.822
Le + Ss + F + Le * Ss	0.013	0.027	2.233	27.591	0.805
W + Le + Ss + F + Ge + Poc + Le * Ss	0.013	0.012	0.924	65.617	0.823
Le + Ss + F + Ge + Le * Ss	0.013	0.005	0.413	145.922	0.807
Le + Ss + F + Poc + Le * Ss	0.013	0.004	0.329	182.965	0.807
W + Le + Ss + F	0.013	6.961×10^{-4}	0.055	1089.774	0.794
Le + Ss + F + Ge + Poc + Le * Ss	0.013	6.672×10^{-4}	0.053	1137.027	0.808
W + Le + Ss + F + Poc	0.013	3.179×10^{-4}	0.025	2386.195	0.799

Note that the prior inclusion probabilities are not equal for all coefficients. This happens because JASP automatically excludes models with interactions effects but without their corresponding main effects, as dictated by the principle of marginality(for details see Nelder, 1977). Thus the prior inclusion probability, $P(\text{incl})$ is still obtained by adding up the prior probability of all models that contain a particular coefficient, but for interaction effects there are simply fewer models that are added up. This is further explained in the section *Including Interaction Effects.*

The change from prior to posterior inclusion probabilities can be visualized

Table 8.4: Model-averaged posterior summary for linear regression coefficients of the Gallup World Happiness Data. The leftmost column denotes the predictor (abbreviations are shown in Table 8.2). The columns 'mean' and 'sd' represent the respective posterior mean and standard deviation of the parameter after model averaging. $P\,(\text{incl})$ denotes the prior inclusion probability and $P\,(\text{incl} \mid \text{data})$ denotes the posterior inclusion probability. The change from prior to posterior inclusion odds is given by the inclusion Bayes factor (BF_{incl}). The last two columns represent a 95% central credible interval (CI) for the parameters.

						95% CI	
Coefficient	Mean	SD	$P\,(\text{incl})$	$P\,(\text{incl}\|\mathcal{D})$	BF_{incl}	Lower	Upper
Intercept	5.346	0.041	1.000	1.000	1.000	5.265	5.421
W	0.263	0.094	0.500	0.962	25.616	0.000	0.393
Le	−0.110	0.035	0.600	1.000	2875	−0.183	−0.050
Ss	−8.545	2.556	0.600	1.000	131 213	−13.688	−4.167
F	1.699	0.345	0.500	1.000	3772	1.067	2.327
Ge	0.028	0.127	0.500	0.115	0.130	−0.037	0.390
Poc	−0.022	0.112	0.500	0.110	0.124	−0.306	0.043
Le * Ss	0.189	0.044	0.200	0.998	2475	0.105	0.267

by selecting `Plots` and ticking `Inclusion probabilities`, which produces the bar graph shown in Figure 8.8.

In addition to providing the inclusion probabilities, Table 8.4 also summarizes the model-averaged posterior distributions using four statistics (i.e., mean, sd, and the lower and upper values of an x% central credible interval). The complete model-averaged posteriors can be visualized by selecting `Plots` and ticking `Marginal posterior distributions`. For example, the posterior distribution for the regression coefficient of Wealth is shown in the left panel of Figure 8.9. The right panel of Figure 8.9 shows the model-averaged posterior for the regression coefficient of Generosity; the spike at zero corresponds to the absence of an effect, and its height reflects the predictor's posterior exclusion probability. The horizontal bar above the distribution shows the 95% central credible interval.

To summarize, the Bayesian model-averaged analysis showed that the most important predictors in the Gallup World Happiness Data are Wealth, Social Support, Life expectancy, and Freedom. There is weak evidence that Generosity and Perception of Corruption are not relevant for predicting Happiness.

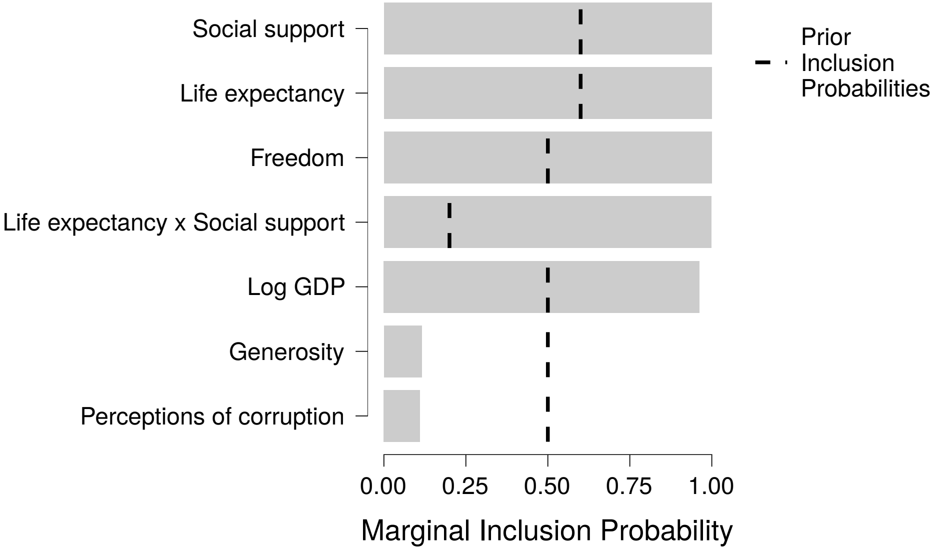


Figure 8.8: Bar graph of posterior inclusion probabilities for the Bayesian linear regression of the Gallup World Happiness Data. The dashed line represents the prior inclusion probabilities. Figure from JASP.

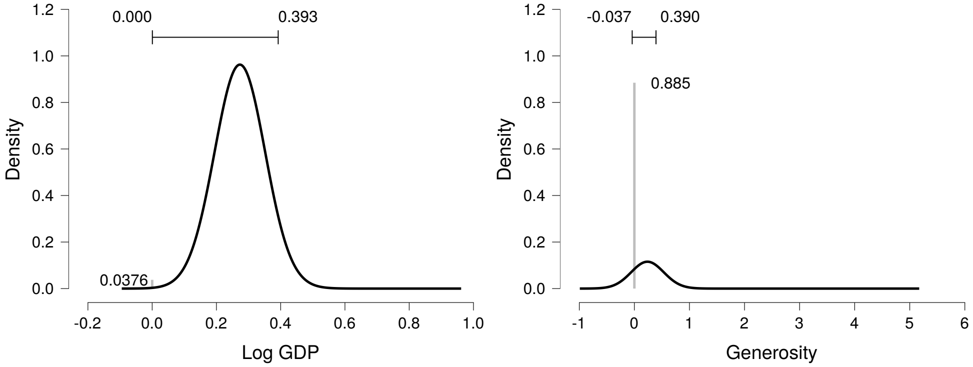
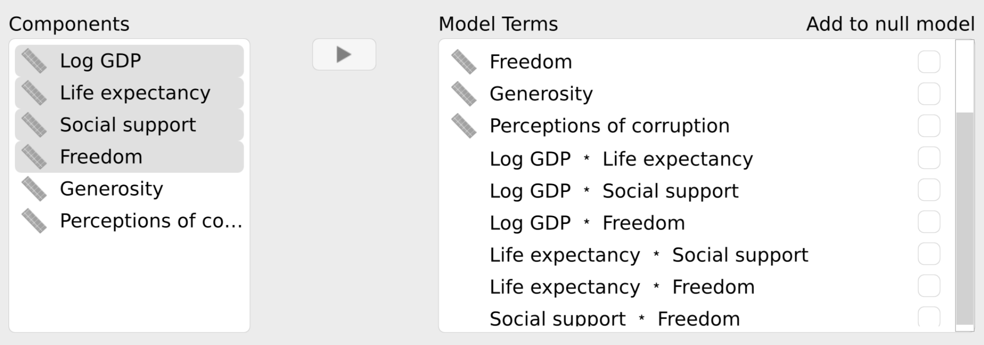


Figure 8.9: The model-averaged posterior of Wealth expressed in GDP (left) and Generosity (right). In the left panel, the number in the bottom left represents the posterior exclusion probability. In the right panel, the posterior exclusion probability is much larger. In both panels, the horizontal bar on top represents the 95% central credible interval. Figures from JASP.

8.2.1 Including Interaction Effects

In regression analysis we are often not interested solely in the main effects of the predictors, but also in the interaction effects. For instance, suppose that for the analysis of the Gallup World Happiness Data we wish to consider the two-way interactions between Wealth, Social Support, Freedom, and Life Expectancy. To do this we click on `Model` and select all variables of interest under `Components` (use ctrl/ ⌘ or Shift to select multiple variables) and drag them to `Model terms`. `JASP` then automatically includes all possible interactions between the selected variables in the `Model terms` on the right. To exclude higher order interactions, we select these in `Model terms` and click the arrow or drag them to `Components`. The result is shown in Figure 8.10.



Components

- Log GDP
- Life expectancy
- Social support
- Freedom
- Generosity
- Perceptions of co...

Model Terms	Add to null model
Freedom	☐
Generosity	☐
Perceptions of corruption	☐
Log GDP * Life expectancy	☐
Log GDP * Social support	☐
Log GDP * Freedom	☐
Life expectancy * Social support	☐
Life expectancy * Freedom	☐
Social support * Freedom	☐

Figure 8.10: Model component view. By selecting multiple variables in the left panel and dragging these to the right panel, all interactions between the selected variables are included in the model. By ticking the box 'Add to null model' the associated variable is included in all models.

As soon as the interaction effects are added to the model, `JASP` updates the output.[6] Since the interaction effects account for 6 new predictors there are now 12 predictors in total and 468 models to consider. There are not $2^{12} = 4096$ models, because JASP automatically excludes models with interactions effects but without their corresponding main effects, as dictated by the principle of marginality (Nelder, 1977). The updated posterior summary is shown in Table 8.5.

Table 8.5 shows that Wealth, Social Support, Life expectancy, and Freedom are important for predicting Happiness, as indicated by the posterior inclusions probabilities. For almost all interaction effects, the posterior inclusion probabilities are smaller than the prior inclusion probabilities, indicating that the data provide evidence against these effects. The interaction effect between Life Expectancy and Social Support somewhat improves the model ($\text{BF}_{\text{incl}} = 8.612$).

[6] When adjusting the model terms it can be inconvenient that JASP continually updates the results. A trick to disable this is to temporarily remove the dependent variable while adjusting the model terms.

Table 8.5: Model-averaged posterior summary for linear regression coefficients of the Gallup World Happiness Data, including two-way interaction effects between Wealth, Social Support, Freedom, and Life Expectancy.

						95% CI	
Coefficient	Mean	SD	$P\,(\text{incl})$	$P\,(\text{incl}\vert\mathcal{D})$	BF_{incl}	Lower	Upper
Intercept	5.346	0.041	1.000	1.000	1.000	5.260	5.425
W	0.233	0.599	0.841	0.982	10.490	−0.945	1.753
Le	−0.122	0.084	0.841	0.997	54.237	−0.288	0.051
Ss	−6.576	4.190	0.841	1.000	3057.789	−12.821	3.223
F	−0.469	2.901	0.841	1.000	1695.479	−6.258	2.608
Ge	0.021	0.117	0.500	0.110	0.124	−0.136	0.236
Poc	−0.015	0.108	0.500	0.106	0.119	−0.409	0.058
W * Le	0.002	0.006	0.363	0.200	0.438	−0.0002	0.019
W * Ss	−0.186	0.599	0.363	0.241	0.557	−1.969	0.660
W * F	0.076	0.237	0.363	0.181	0.389	−0.066	0.788
Le * Ss	0.168	0.116	0.363	0.831	8.612	0.000	0.402
Le * F	0.011	0.035	0.363	0.180	0.385	−0.0001	0.117
Ss * F	1.072	2.562	0.363	0.228	0.517	−0.263	8.086

Comparing the main effects in Table 8.4 to those in Table 8.5, it might appear surprising that the support for including the predictors decreased for all variables. For example, the inclusion Bayes factor for Life Expectancy decreased from about 2875 to 54, Wealth decreased from about 26 to 10, and the interaction between Life Expectancy and Social support decreased from about 2475 to 9. The cause for these change lies in the added interaction effects. All interaction effects with Wealth led to poorly performing models, as illustrated by the low inclusion Bayes factors for all interaction effects with Wealth. As a consequence, the inclusion Bayes factor for Wealth also suffered, since 312 out of the 396 models considered to calculate the inclusion Bayes factor contained interaction effects with Wealth.

The effect of model averaging on parameter estimation is clearly present when comparing the 95% credible intervals in Tables 8.4 and 8.5. For instance, the credible interval for Freedom was [1.06, 2.35] in Table 8.4 but widens to [−6.3, 2.6] in Table 8.5. There are two reasons for this increase in uncertainty. First, the posterior probability of the best model is only 0.223, compared to 0.759 in Table 8.3 (see the online supplement for all posterior model probabilities). This means that other models contribute substantially to the model-averaged posterior, which increases the uncertainty in the parameter estimates. Second, the results in Table 8.5 are based on a larger model space, which potentially leads to a wider range of possible estimates and hence increases the associated uncertainty.

The instability of the results due to changing the model space is no reason for concern; rather, it demonstrates the importance of considering all models and dealing with model uncertainty appropriately. The example above does show, however, that some rationale should be provided for the model space. Here, we did not properly motivate the inclusion of the interaction effects because we wanted to demonstrate the effect of model uncertainty on the results. Instead, one should decide upon the the model space before executing the analysis and ideally preregister the model space on the basis of substantive considerations.

8.2.2 Including Nuisance Predictors in All Models

Another common procedure in the toolkit of linear regression is to include a number of nuisance predictors in all models (in management sience this is sometimes called hierarchical regression; see also Andraszewicz et al., 2015; Petrocelli, 2003). Subsequently, the goal is to assess the contribution of the predictor(s) of interest over and above the contribution from the nuisance predictors. For example, we could have included Wealth in all models, for instance because we already know that Wealth has a large effect, but we are not interested in that effect – we are interested in what the other predictors add on top of Wealth. To add Wealth as a nuisance variable to the model, we go to `Model` and check the box under `Add to null model` for Wealth (see Figure 8.10). As with interaction effects, `JASP` updates the results immediately and produces a model comparison table similar to Table 8.3. Note that the Bayes factor BF_{01} in the fifth column of Table 8.3 by default compares all models to the *best* model. When including nuisance predictors, we are more interested in how much the models improve compared to the null model. We can change the default setting by going to `Order` and selecting `Compare to null model`. This changes the Bayes factor column such that all models are compared to the null model instead of to the best model. The resulting table is shown in Table 8.6. Since we now compare all models to the null model, the null model is always shown in the first row.

8.2.3 Prior Sensitivity

Priors on Parameters

In the previous analyses we used the default JZS prior on the values of the regression coefficients. However, it is generally recommended to investigate the robustness of the results against the choice of prior (van Doorn et al., 2020). To investigate robustness, one typically uses the same family of distributions but varies the prior width. A wider prior will imply more spread-out a-priori

Table 8.6: The 10 best models from the Bayesian linear regression for the Gallup World Happiness Data, where the nuisance predictor Wealth is included in all models. The interpretation of the columns is identical to that of Table 8.3, except that the Bayes factor BF_{01} in the fifth column compares all models to the null model. The table footnote shows a reminder from `JASP` which variables are specified as nuisance.

Models	$P(\mathcal{M})$	$P(\mathcal{M} \mid \text{data})$	$\text{BF}_{\mathcal{M}}$	BF_{01}	R^2
Null model (incl. W)	0.031	6.143×10^{-11}	1.904×10^{-9}	1.000	0.679
Le + Ss + F	0.031	0.439	24.228	7.141×10^{9}	0.794
Le + Ss + F + Poc	0.031	0.200	7.767	3.261×10^{9}	0.799
Le + Ss + F + Ge	0.031	0.169	6.290	2.746×10^{9}	0.799
Ss + F	0.031	0.077	2.572	1.247×10^{9}	0.781
Le + Ss + F + Ge + Poc	0.031	0.043	1.380	6.938×10^{8}	0.802
Ss + F + Poc	0.031	0.032	1.034	5.254×10^{8}	0.786
Ss + F + Ge	0.031	0.030	0.955	4.867×10^{8}	0.786
Ss + F + Ge + Poc	0.031	0.007	0.217	1.131×10^{8}	0.789
Le + F	0.031	0.002	0.057	2.966×10^{7}	0.769

Note. All models include Wealth (W).

uncertainty about the effect, whereas a more narrow prior implies that the a-priori belief about the effect is more concentrated near zero. To adjust the prior, we go to `Advanced options` and under `Prior` change the value after `JZS`. This value is generally referred to as the scale of the JZS prior. The default choice in JASP is a JZS with a scale of 1/8. This corresponds to the default choice used in other software, for example the R package “BayesFactor” (Morey & Rouder, 2021). If the JZS scale in `JASP` is s, the corresponding scale for the “BayesFactor” package is $\sqrt{2s}$. Commonly used values for the larger scales are 1/4 and 1/2, respectively referred to as “wide” and “ultrawide” priors (Morey & Rouder, 2021; Wagenmakers, Love, et al., 2018). Figure 8.11 shows the marginal prior distribution for the regression coefficients β for these three scales. Under `Advanced options` it is also possible to select other prior distributions than the JZS. However, we recommend against doing so without proper motivation (see e.g., Bayarri et al., 2012; Consonni et al., 2018; Liang et al., 2008).

We repeated the main analysis with a JZS scale of 1/4 and 1/2 but the posterior inclusion probabilities, see Table 8.7, did not change in a meaningful way (see https://osf.io/5dmj7/ for an annotated `.jasp` file with the results).

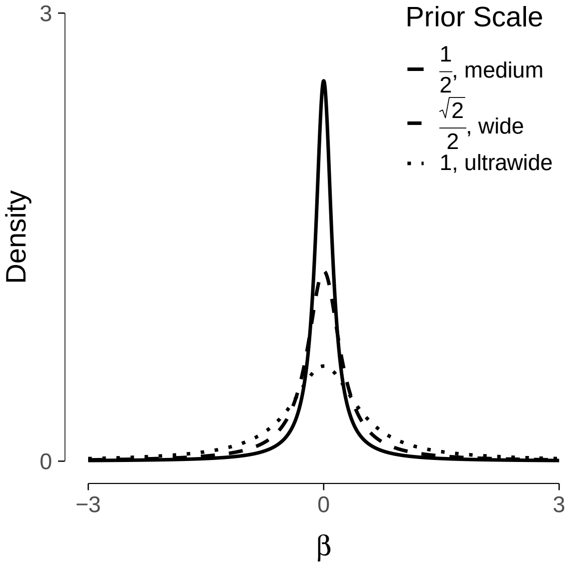


Figure 8.11: Marginal prior distribution on the regression coefficients (β). The different line types represent different scales for the prior. As the scale increases the probability mass near zero decreases and the mass on more extreme values increases.

Table 8.7: Posterior inclusion probabilities given different values for the scale of the JZS prior. The intercept is omitted from the comparison as it is included in all models and therefore its inclusion probability is always 1.

		$P\,(\text{incl}\|\mathcal{D})$		
Coefficient	P(incl)	s = medium	s = wide	s = ultrawide
Log GDP	0.5	0.962	0.962	0.962
Le	0.6	1.000	1.000	1.000
Ss	0.6	1.000	1.000	1.000
F	0.5	1.000	1.000	1.000
G	0.5	0.115	0.114	0.111
Poc	0.5	0.110	0.109	0.106
Le * Ss	0.2	0.998	0.998	0.998

Priors on the Model Space

Aside from adjusting the priors on the coefficients, it is also possible to adjust the prior over the models. An intuitive choice is a uniform model prior, where each model is assigned prior mass equal to one over the number of models considered. This prior was also used in the analyses above. However, if we use a uniform model prior and then compute the prior probability for a model that includes x predictors, where x goes from 0 to p, we do not obtain a uniform prior. Instead, the implied prior over the number of included predictors is bell-shaped with the most mass on models with $p/2$ predictors. Thus, a-priori our prior is biased against sparse models and dense models, and favors something in between.

A solution to this problem is to use a prior that is uniform over the number of included predictors. This can be achieved by dividing the total probability, 1, into $p+1$ chunks. The first chunk represents the combined probability of all models that include no predictors, the second chunk represents the combined probability of all models that include one predictor, etc. This model prior commonly referred to as a beta-binomial model prior and can be tweaked using two parameters, α and β. The left panel of Figure 8.12 shows how the total probability is divided for different values of α and β. The default values in JASP are $\alpha = \beta = 1$.[7] In the next step, all models within a chunk (i.e. all models with the same number of predictors) are treated as equally likely and the probability of the chunk is distributed uniformly among them. This implies the prior probability of a chunk is divided by the number of models in that chunk. The right panel of Figure 8.12 shows the prior model probability for different values of α and β.

We repeated the main analysis with a Beta-binomial prior. Table 8.8 shows the inclusion probabilities for an uniform model prior and a beta-binomial model prior. Although the numbers differ, the results are unchanged: The evidence for the inclusion and exclusion of predictors in the model point in the same direction for both priors on the model space. For example, the inclusion Bayes factors that were larger than 1 for a uniform prior on the model space were also larger than 1 for the beta-binomial prior.

Although much attention goes to the choice of prior distribution, the likelihood of the statistical model is often more important. As stated by Gelman and Robert (2013):

> "*It is perhaps merely an accident of history that skeptics and subjectivists alike strain on the gnat of the prior distribution while swallowing the camel that is the likelihood.* " (italics in original)

[7] The α and β parameters of the beta-binomial prior can be set individually. Alternatively it is possible to choose the Wilson model prior or the Castillo model prior, which are both variants of the beta-binomial prior (Castillo et al., 2015; M. A. Wilson et al., 2010). The Wilson model prior sets $\alpha = 1$ and $\beta = \lambda p$, where p is the number of predictors in the model and λ is a parameter set by the user. The Castillo model prior sets $\alpha = 1$ and $\beta = p^u$, where p is the number of predictors in the model and u is a parameter set by the user. Both the Wilson and the Castillo prior assign more mass to models with fewer predictors.

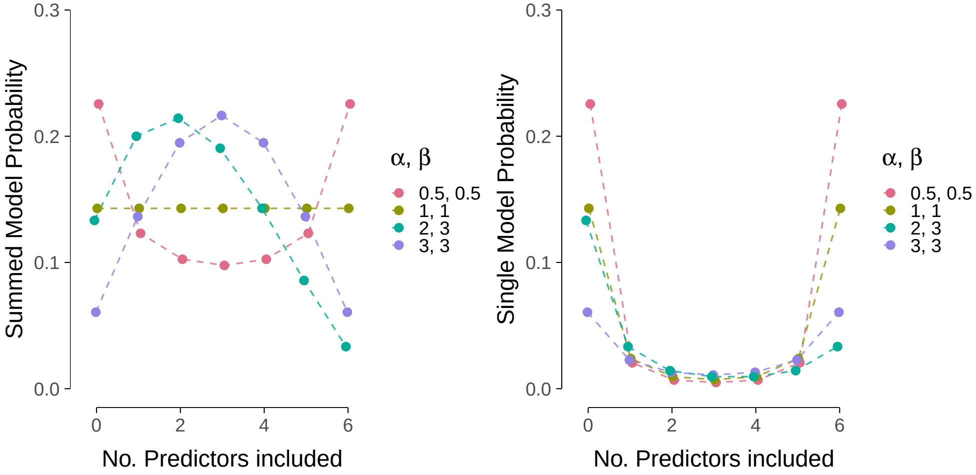


Figure 8.12: A beta-binomial model prior for a model space with 6 predictors. The left panel shows the beta-binomial distribution where the number of predictors in the model (x-axis) is visualized against the total probability of all models with that number of predictors (y-axis). The right panel shows how the number of predictors in the model (x-axis) influences the prior probability of a single model (y-axis). The right panel is obtained by dividing each probability in the left panel by the number of models with that many predictors. The number of models that contain j predictors is obtained by calculating $\binom{6}{j}$. This yields for 0 through 6: 1, 6, 15, 20, 15, 6, and 1.

Table 8.8: Prior inclusion probabilities, posterior inclusion probabilities, and inclusion Bayes factors for a uniform model prior and a beta-binomial model prior. The intercept is omitted from the comparison as it is included in all models and therefore its inclusion probability is always 1.

	Uniform			Beta-binomial		
Coefficient	$P\,(\text{incl})$	$P\,(\text{incl}\mid\mathcal{D})$	BF_{incl}	$P\,(\text{incl})$	$P\,(\text{incl}\mid\mathcal{D})$	BF_{incl}
Log GDP	0.5	0.962	25.616	0.489	0.983	59.024
Le	0.6	1.000	2875	0.556	1.000	8924
Ss	0.6	1.000	131213	0.556	1.000	398502
F	0.5	1.000	3772	0.489	1.000	5775
G	0.5	0.115	0.130	0.489	0.339	0.536
Poc	0.5	0.110	0.124	0.489	0.330	0.515
Le * Ss	0.2	0.998	2475	0.333	0.999	2336

In other words, choices about which predictors and interaction effects to consider, choices that influence the likelihood, are more important than the choice of prior distribution. This again stresses the importance to demarcate the model space.

8.3 Discussion

This paper provided a tutorial on Bayesian multi-model inference and aimed to bridge the gap between statistical theory and the applied researcher. Multi-model inference and regression analyses are subject to a number of limitations, which are discussed below.

8.3.1 Limitations

At the moment of writing, the linear regression procedures as implemented in `JASP` and BAS do not account for missing values; therefore, missing values are deleted list-wise (i.e., cases with missing values for one or more predictors are omitted from the analysis entirely). However, Bayesian analyses can handle missing values by perceiving them as unknown parameters of the model. That way, the observed value can still contribute to the model and the uncertainty around the missing values is dealt with accordingly (Little & Rubin, 2002, Ch 10).

A general challenge for regression models arises when the predictors are multicollinear, that is, very highly correlated. To illustrate, consider the data of 13 American football punters (available from Faraway, 2005). The goal is to relate various physical characteristics of the football players to their average punting distance. Relevant predictors are right leg strength, left leg strength, right hamstring flexibility, and left hamstring flexibility. Unsurprisingly, the correlation between the right and left leg predictors is very high. Consequently, models that contain predictors from one leg benefit little when the predictor from the other leg is added on top. Thus, models with predictors for both legs perform poorly compared to models containing information of only one leg. After calculating the inclusion Bayes factors it is unclear whether any specific predictor should be included. Paradoxically, when directly comparing the models, the null model is one of the worst models; it performs about 31.8 times worse than the best model with right hamstring flexibility as the only predictor. See `punting.jasp` at https://osf.io/5dmj7/ for an annotated analysis. Nonetheless, these results make sense. The model averaged results are unable to distinguish between the correlated predictors because individually they improve the model but jointly they worsen it. For example, the second best model contains right leg strength as a predictor, the fifth best model contains left leg strength as a predictor, but the model that contains both right and left leg strength as predictors ranks 11th out of 16. Hence, there is a lingering uncertainty about which predictor to include, even though directly comparing the different models shows that a model including at least one predictor already performs better than the null model.

Recognizing multicollinearity is always important in linear regression. This does not require much additional work; when creating Figure 8.4, the pairwise correlations can also be examined. Another way to assess multicollinearity is by calculating the variance inflation factor (Sheather, 2009, Ch. 6.4).

8.3.2 Violation of Assumptions

If the assumption of linearity appears violated for one or more predictors, some transformations can be used (e.g., a log-transformation). Alternatively, one could try including the square (or cube) of a predictor, and including that in the regression equation to capture any nonlinear relations. This is also known as polynomial regression and can be used to relax the linearity assumption. In `JASP`, polynomial regression or other transformations can be managed easily using `Compute Columns`. If the relation between the criterion variable and predictors is innately non-linear, for instance because the criterion variable is binary, generalized linear models can be used. The R package BAS can also be used for multi-model inference for generalized linear models.

If the residuals appear non-normal or heteroscedastic, then there is no clear way how to proceed. Ideally, one first identifies the cause of the violation. Violations can be caused by a single predictor with a nonlinear relation causing misfit, or by multiple predictors. Nonlinearities can be dealt with using the suggestions in the previous paragraph. If the source remains unclear, or is innate to the data, alternative methods can be used. One alternative is to use a probabilistic programming language suited for general Bayesian inference, such as JAGS (Plummer, 2003), NIMBLE (de Valpine et al., 2017), OpenBUGS (Lunn et al., 2009), or MultiBUGS (Goudie et al., 2017), all of which are conceptual descendants of WinBUGS (Lunn et al., 2000; Ntzoufras, 2009). The main advantage of probabilistic programming languages is their flexibility: for instance, models can be adjusted to accommodate heteroscedastic residuals (e.g., Reich & Ghosh, 2019, Ch. 4.5.2). These languages also come with disadvantages. First, it is easier to make a mistake – either a programming error, a statistical error, or both. Second, the languages are generic, and because they are not tailored to specific applications they may be relatively inefficient compared to a problem-specific method.

In sum, the goal of this tutorial was to familiarize applied researchers with the theory and practice of Bayesian multi-model inference. By accounting for model uncertainty in regression it is possible to prevent the overconfidence that inevitable arises when all inference is based on a single model. We hope that tutorial will enable applied researchers to use Bayesian multi-model inference in their own work.

8.3.3 Acknowledgements

Annotated `.jasp` files and supplementary materials can be found at https://osf.io/5dmj7/.

9

A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP

Analysis of variance (ANOVA) is the standard procedure for statistical inference in factorial designs. Typically, ANOVAs are executed using frequentist statistics, where *p*-values determine statistical significance in an all-or-none fashion. In recent years, the Bayesian approach to statistics is increasingly viewed as a legitimate alternative to the *p*-value. However, the broad adoption of Bayesian statistics –and Bayesian ANOVA in particular– is frustrated by the fact that Bayesian concepts are rarely taught in applied statistics courses. Consequently, practitioners may be unsure how to conduct a Bayesian ANOVA and interpret the results. Here we provide a guide for executing and interpreting a Bayesian ANOVA with JASP, an open-source statistical software program with a graphical user interface. We explain the key concepts of the Bayesian ANOVA using two empirical examples.

This chapter is published as: van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology.*, *120*(1), 73–96.

Ubiquitous across the empirical sciences, analysis of variance (ANOVA) allows researchers to assess the effects of categorical predictors on a continuous outcome variable. Consider for instance an experiment by Strack et al. (1988) designed to test the *facial feedback hypothesis*, that is, the hypothesis that people's affective responses can be influenced by their own facial expression. Participants were randomly assigned to one of three conditions. In the *lips* condition, participants were instructed to hold a pen with their lips, inducing a pout. In the *teeth* condition, participants were instructed to hold a pen between their teeth, inducing a smile. In the control condition, participants were told to hold a pen in their nondominant hand. With the pen in the instructed position, each participant then rated four cartoons for funniness. The outcome variable was the average funniness rating across the four cartoons. The ANOVA procedure may be used to test the null hypothesis that the pen position does not result in different funniness ratings.

ANOVAs are typically conducted using frequentist statistics, where p-values decide statistical significance in an all-or-none manner: if $p < .05$, the result is deemed statistically significant and the null hypothesis is rejected; if $p > .05$, the result is deemed statistically nonsignificant, and the null hypothesis is retained. Such binary thinking has been critiqued extensively (e.g., Amrhein et al., 2019; Cohen, 1994; Rouder et al., 2016), and some perceive it as a cause of the reproducibility crisis in psychology (Cumming, 2014; but see Savalei and Dunn, 2015). In recent years, several alternatives to p-values have been suggested, for example reporting confidence intervals (Cumming, 2014; Gardner & Altman, 1986) or abandoning null hypothesis testing altogether (McShane et al., 2019).

Here we focus on another alternative: Bayesian inference. In the Bayesian framework, knowledge about parameters and hypotheses is updated as a function of predictive success – hypotheses that predicted the observed data relatively well receive a boost in credibility, whereas hypotheses that predicted the data relatively poorly suffer a decline (Wagenmakers et al., 2016). A series of recent articles show how the Bayesian framework can supplement or supplant the frequentist p-value (e.g., Burton et al., 1998; Dienes & McLatchie, 2018; Jarosz & Wiley, 2014; Masson, 2011; Nathoo & Masson, 2016; Rouder et al., 2016).

The advantages of the Bayesian paradigm over the frequentist p-value are well documented (e.g., Wagenmakers, Marsman, et al., 2018); for instance, with Bayesian inference researchers can incorporate prior knowledge and quantify support, both in favor and against the null-hypothesis; furthermore, this support may be monitored as the data accumulate (Stefan et al., 2019). Despite these and other advantages, Bayesian analyses are still used only sparingly in the social sciences (van der Schoot et al., 2017). The broad adoption of Bayesian statistics –and Bayesian ANOVA in particular– is hindered by the fact that Bayesian concepts are rarely taught in applied statistics courses. Consequently, practitioners may be unsure of how to conduct a Bayesian ANOVA and interpret the results.

To help familiarize researchers with Bayesian inference for common experimental designs, this article provides a guide for conducting and interpreting a Bayesian ANOVA with JASP (JASP Team, 2022). JASP is a free, open-source statistical software program with a graphical user interface that offers both Bayesian and frequentist analyses. Below, we first provide a brief introduction to Bayesian statistics. Subsequently, we use two data examples to explain the key concepts of ANOVA.

9.1 Bayesian Foundations

This section explains some of the fundamentals of Bayesian inference. We focus on interpretation rather than mathematical detail; see the special issue on Bayesian inference by (Vandekerckhove et al., 2018) for a set of comprehensive, low-level introductions to Bayesian inference.

The central goal of Bayesian inference is learning, that is, using observations to update knowledge. In an ANOVA we want to learn about the candidate models $\boldsymbol{\mathcal{M}}$ and their condition-effect parameters $\boldsymbol{\beta}$. Returning to the example of the facial feedback experiment, we commonly specify two models. The null model describes the funniness ratings using a single grand average across all three conditions, effectively stating that there is no effect of pen position. The parameters of the null model are thus the average test score and the error variance. The alternative model describes the funniness ratings using an overall average and the effect of pen position; in other words, the means of the three condition are allowed to differ. Therefore, the alternative model has five parameters: the average funniness ratings across participants, the error variance, and for each of the three pen positions the magnitude of the effect.[1]

To start the learning process we need to specify prior beliefs about the plausibility of each model, $p(\boldsymbol{\mathcal{M}})$, and about the plausible parameters values $\boldsymbol{\beta}$ within each model, $p(\boldsymbol{\beta} \mid \boldsymbol{\mathcal{M}})$. These prior beliefs are represented by *prior distributions*. Observing data $\mathcal{D}$ drives an update of beliefs, transforming the prior distribution over models and parameters to a joint *posterior distribution*, denoted $p(\boldsymbol{\beta}, \boldsymbol{\mathcal{M}} \mid \mathcal{D})$.[2] The updating factor –the change from prior to posterior beliefs– is determined by relative predictive performance for the observed data

[1] Note that one of the four parameters, average funniness rating and the three condition effects, is redundant. That is, we can make identical predictions even when we fix one of the four parameters to zero.

[2] The joint posterior distribution describes both the marginal probability distribution, for example, $p\boldsymbol{\beta} > 0 \mid \mathcal{D}$ or $p\mathcal{M} \mid \mathcal{D}$, and the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\mathcal{M}}$, for example, in a particular model the posterior mass of $\boldsymbol{\beta}$ is more concentrated around 0.

(Wagenmakers et al., 2016). As shown in Figure 9.1, the knowledge updating process forms a learning cycle, such that the posterior distribution after the first batch of data becomes the prior distribution for the next batch.

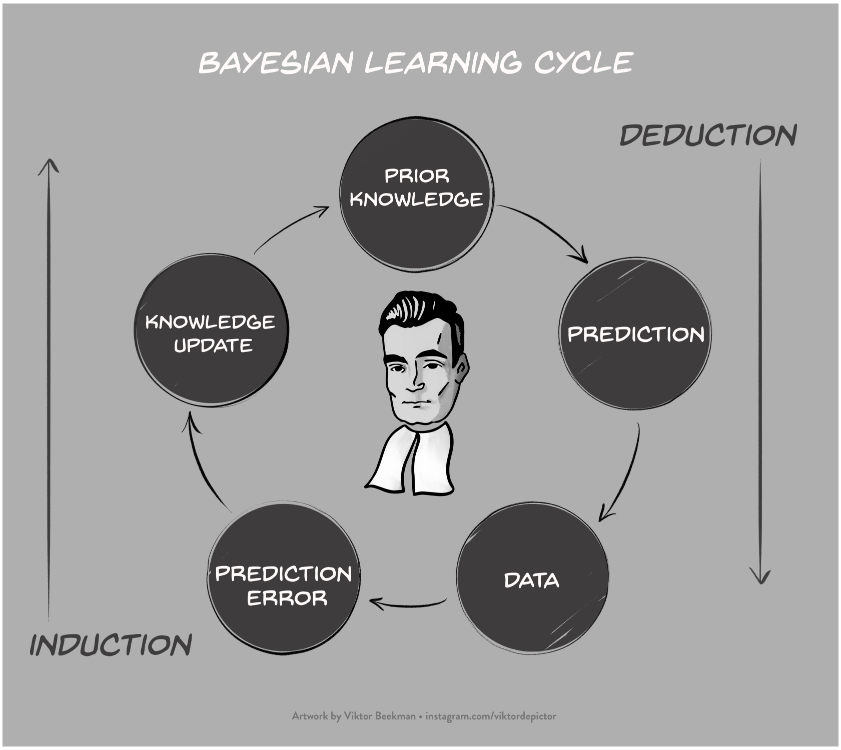


Figure 9.1: Bayesian learning can be conceptualized as a cyclical process of updating knowledge in response to prediction errors. The prediction step is deductive, and the updating step is inductive. For a detailed account see Jevons (Chapters XI and XII 1874/1913). Figure available at BayesianSpectacles.org under a CC-BY license.

Mathematically, the updating process is given by Bayes' rule:

$$\overbrace{p(\boldsymbol{\beta},\boldsymbol{\mathcal{M}}\mid\mathcal{D})}^{\substack{\text{Joint posterior}\\ \text{distribution}}} = \overbrace{p(\boldsymbol{\mathcal{M}})}^{\substack{\text{Prior model}\\ \text{probability}}} \times \overbrace{p(\boldsymbol{\beta}\mid\boldsymbol{\mathcal{M}})}^{\substack{\text{Prior param.}\\ \text{probability}}} \times \overbrace{\frac{p(\mathcal{D}\mid\boldsymbol{\beta},\boldsymbol{\mathcal{M}})}{p(\mathcal{D})}}^{\text{Updating factor}}. \tag{9.1.1}$$

This rule stipulates how knowledge about the relative plausibility of both models and parameters ought to be updated in light of the observed data. When the focus is on the comparison of two rival models, one generally considers only the model updating term. This term, commonly known as the *Bayes factor*, quantifies the relative predictive performance of the rival models, that is, the change in relative model plausibility that is brought about by the data (Etz & Wagenmakers, 2017; Jeffreys, 1939; Kass & Raftery, 1995; Wrinch & Jeffreys, 1921):

$$\underbrace{\frac{p(\mathcal{M}_1 \mid \mathcal{D})}{p(\mathcal{M}_0 \mid \mathcal{D})}}_{\text{Posterior model odds}} = \underbrace{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}}_{\text{Prior model odds}} \times \underbrace{\frac{p(\mathcal{D} \mid \mathcal{M}_1)}{p(\mathcal{D} \mid \mathcal{M}_0)}}_{\substack{\text{Bayes factor}\\ \text{BF}_{10}}}. \tag{9.1.2}$$

When the Bayes factor BF_{10} equals 20, the observed data are twenty times more likely to occur under $\mathcal{M}_1$ than under $\mathcal{M}_0$ (i.e., support for $\mathcal{M}_1$ versus $\mathcal{M}_0$); when the Bayes factor BF_{10} equals 1/20, the observed data are twenty times more likely to occur under $\mathcal{M}_0$ than under $\mathcal{M}_1$ (i.e., support for $\mathcal{M}_0$ versus $\mathcal{M}_1$); when the Bayes factor BF_{10} equals 1, the observed data are equally likely to occur under both models (i.e., neither model is supported over the other). Note that the Bayes factor is a comparison of two models and hence it is always a relative measure of evidence, that is, it quantifies the performance of one model relative to another.[3] Likewise, the prior and posterior odds are both odds ratios, which means they are relative measures of an effect. For example, a posterior odds of 2 means that the model in the numerator is twice as likely as the model in the denominator, after we observe the data. The Bayes factor can be presented as BF_{10}, $p(\mathcal{D}|\mathcal{M}_1)$ divided by $p(\mathcal{D}|\mathcal{M}_0)$, or as its reciprocal BF_{01}, $p(\mathcal{D}|\mathcal{M}_0)$ over $p(\mathcal{D}|\mathcal{M}_1)$. Typically, BF_{10} is used to present evidence in favor of the alternative hypothesis whereas BF_{01} is used to present evidence in favor of the null hypothesis.

The Bayesian paradigm differs from the frequentist paradigm in at least three key aspects. First, evidence in favor of a particular model, quantified by a Bayes factor, is a *continuous* measure of support. Unlike the frequentist Neyman-Pearson decision rule (usually $p < 0.05$), there is no need to impose all-or-none Bayes factor cut-offs for accepting or rejecting a particular model. Moreover, the Bayes factor can discriminate between "absence of evidence" (i.e., nondiagnostic data that are predicted about equally well under both models, such that the Bayes factor is close to 1) and "evidence of absence" (i.e., diagnostic data that support the null hypothesis over the alternative hypothesis).

A second difference is that, in the Bayesian paradigm, knowledge about models $\boldsymbol{\mathcal{M}}$ and parameters $\boldsymbol{\beta}$ is updated simultaneously. Consequently, it is natural to account for model uncertainty by considering all models, but assigning more weight to those models that predicted the data relatively well. This procedure is known as Bayesian model averaging (BMA; Hinne et al., 2020;

[3] For a cartoon that explains the strength of evidence provided by a Bayes factor, see https://www.bayesianspectacles.org/lets-poke-a-pizza-a-new-cartoon-to-explain-the-strength-of-evidence-in-a-bayes-factor/

Hoeting et al., 1999; Jevons, 1874/1913; Jeffreys, 1939, p. 296; Jeffreys, 1961, p. 365). In contrast, many frequentist analyses first select a ‘best’ model and subsequently estimate its parameters, thereby neglecting model uncertainty and producing overconfident conclusions (Ch 7.4 Claeskens & Hjort, 2008).[4] Another benefit of BMA is that point estimates and uncertainty intervals can be derived without conditioning on a specific model. This way, model uncertainty is accounted for in point estimates and uncertainty intervals.

A third difference is that the Bayesian posterior distributions allow for direct probabilistic statements about parameters. For example, based on the posterior distribution of $\boldsymbol{\beta}$ we can state that we are 95% confident that the parameter lies between x and y. This range of parameter values is commonly known as a *95% credible interval.*[5] Similarly, we can consider any interval from a to b and quantify our confidence that the parameter falls in that specific range.

A fourth difference is that Bayesian inference automatically penalizes for complexity and thus favors sparsity. Consider a model with a redundant covariate. Since the covariate is redundant, a model with this covariate will make poor predictions. Consequently, the Bayes factor, which compares the relative predictive performance of two models, will favor the model without the redundant predictor over the model with the redundant predictor. Key is that the predictive performance is assessed using parameter values that are drawn from the prior distributions.

9.2 ANOVA

Traditionally, analysis of variance involves –as the name suggests– a comparison of variances. In the frequentist framework, the variance between each level of the categorical predictor is compared to the variance within the levels of the categorical predictor.

When the categorical predictor has no effect, the population variances between the levels equals the population variances within the levels, and the sample ratio of these variances is distributed according to a central F-distribution. Under the assumption that the null hypothesis is true, we may then calculate the probability of encountering a sample ratio of variances that is at least as large as the one observed – this then yields the much-maligned yet omnipresent

[4] Although uncommon, it is possible to average over the models in the frequentist framework. To do so, calculate for each model an information criterion such as AIC, and use a transformed version as model weights (Burnham & Anderson, 2002).

[5] Note the difference in interpretation compared to the frequentist 95% confidence interval: “if we repeat this experiment an infinite number of times and compute an infinite number of confidence intervals, then 95% of these intervals contain the true parameter value.” See also Morey, Hoekstra, et al. (2016).

p-value.

Instead, the Bayesian ANOVA contrasts the predictive performance of competing models (Rouder et al., 2016). In order to make predictions the model parameters need to be assigned prior distributions. These prior distributions could in principle be specified from subjective background knowledge, but here we follow Rouder et al. (2012) and use a default specification inspired by linear regression models, designed to meet general desiderata such as consistency and scale invariance (i.e., it does not matter whether the outcome variable is measured in seconds or milliseconds; see also Bayarri et al., 2012; Liang et al., 2008).

9.3 Assumptions

Before interpreting the results from an ANOVA, it is prudent to assess whether its main assumption holds, namely that the residuals are normally distributed. A common tool to assess the normality of the residuals is a Q-Q plot, which visualizes the quantiles of the observed residuals against the quantiles expected from a standard normal distribution. If the residuals are normally distributed then all the points in a Q-Q plot fall on the red line in Figure 9.2. In contrast to a frequentist ANOVA, where the residuals are point estimates, a Bayesian ANOVA provides a probability distribution for each residual. The uncertainty in the residuals can thus be summarized by 95% credible intervals. The left panel of Figure 9.2 shows an example where the larger quantiles lie away from the red line, displaying a substantial deviation from normality. The right panel of Figure 9.2 shows residuals that are more consistent with what is expected under a normal distribution.

Introductory texts discuss additional ANOVA assumptions, most of which follow directly from the normality of the residuals. For some of these assumptions, violations can be difficult to detect visually in a Q-Q plot. An example is sphericity, which is specific to repeated measures ANOVA. One definition of sphericity is that the variance of all pairwise difference scores is equal. In the frequentist paradigm, this assumption is usually assessed using Mauchly's test (but see Tijmstra, 2018). Another example is homogeneity of variances, which implies that the residual variance is equal across all levels of the predictors. Homogeneity of variances can be assessed using Levene's test (Levene, 1961).

The following sections illustrate how to conduct and interpret a Bayesian ANOVA with JASP. JASP can be freely downloaded from https://jasp-stats.org/download/. Annotated `.jasp` files of the discussed analyses, data sets, and a step-by-step guide on conducting a Bayesian ANOVA in JASP are available at https://osf.io/f8krs/. We should stress that the current implementation of Bayesian ANOVA in JASP is based on the R package *BayesFactor*

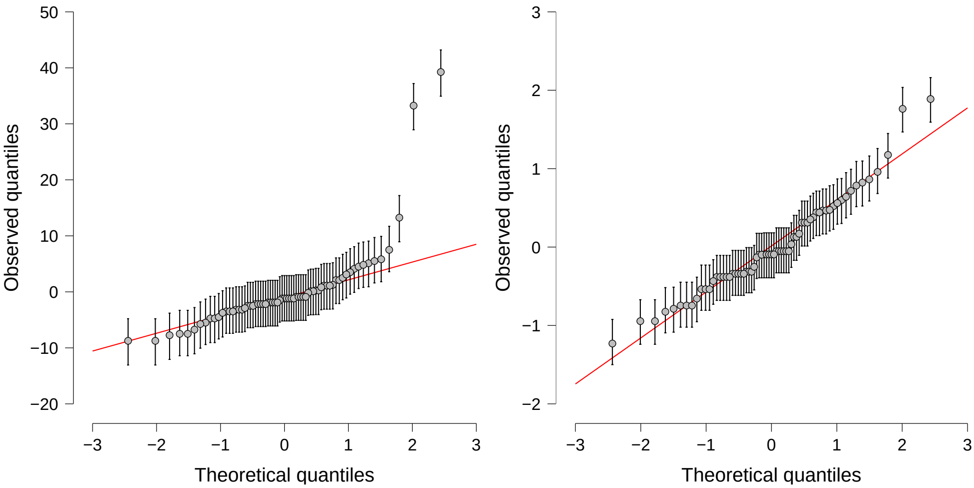


Figure 9.2: Q-Q plots of non-normally distributed residuals (left) and approximately normally distributed residuals (right). The vertical bars through each point represent the 95% central credible interval. If the data are perfectly normally distributed, all points fall on the red line. Note that the y-axis of the two panels has a different scale.

(Morey & Rouder, 2021) which is itself based on the statistical work by Rouder et al. (2012).

9.4 Example I: A Robot's Social Skills

Do people take longer to switch off a robot when it displays social skills? This question was studied by Horstmann et al. (2018) and we use their data to illustrate the key concepts of a Bayesian ANOVA. In the Horstmann et al. (2018) study, 85 participants interacted with a robot. Participants were told that the purpose of their interaction with the robot was to test a new algorithm. After two dummy tasks were completed, the instructor told the participants that they could switch off the robot if they wanted. The outcome variable was the time it took participants to switch off the robot. Here we analyze the log-transformed switch-off times since the Q-Q plot of the raw switch off times showed a violation of normality. Horstmann et al. (2018) manipulated two variables in a between-subjects design. First, they manipulated the robots' verbal responses to be either social (e.g., "Oh yes, pizza is great. One time I ate a pizza as big as me.") or functional (e.g., "You prefer pizza. This worked well. Let us continue."). Second, either the robot protested to being turned off (e.g., "No! Please do not switch me off! I am scared that it [*sic*] will not brighten up again!") or it did not. Therefore, the design of this study is a 2x2

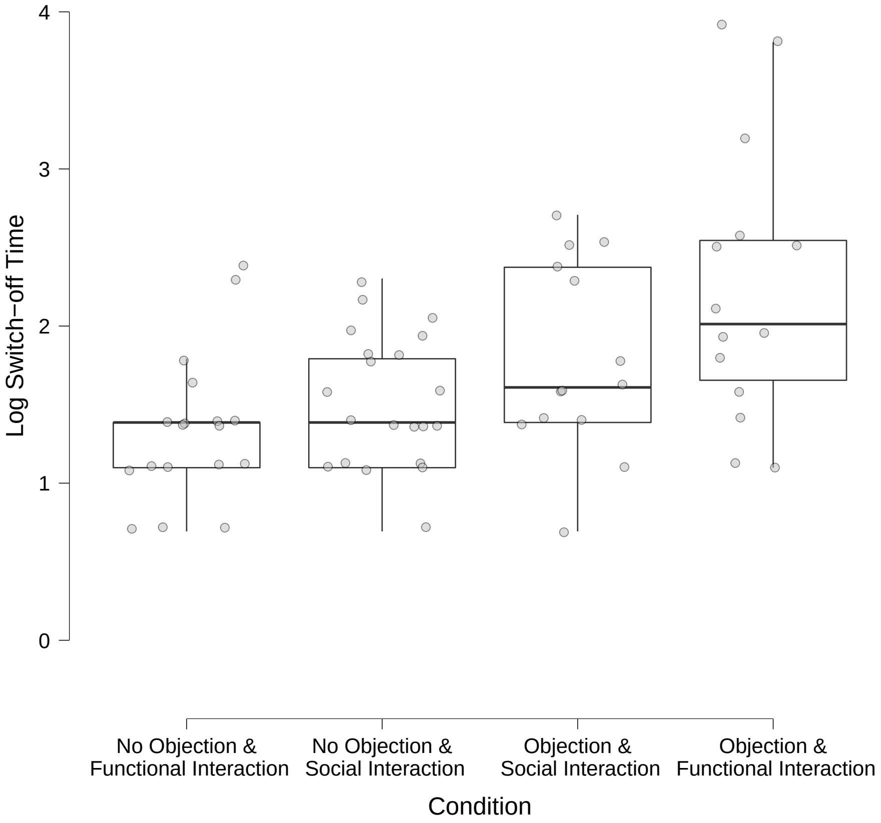


Figure 9.3: Observed log switch-off times for the data of Horstmann et al. (2018).

between-subjects ANOVA. The data are shown in Figure 9.3.

9.4.1 Interpreting the Bayesian ANOVA

Model comparison The primary output from the JASP ANOVA is presented in Table 9.1, which shows the support that the data offer for each model under consideration. The left-most column lists all models at hand: four alternative models and one null model. The models are ordered by their predictive performance relative to the best model; this is indicated in the BF_{01} column, which shows the Bayes factor relative to the best model which features only the objection factor. For example, the data are about 73 times more likely under the model with only the robot's objection as a predictor than under the null model. The prior model probability $P(\mathcal{M})$ is 0.2 for all models and the resulting posterior model probabilities are given by $P(\mathcal{M} \mid \mathcal{D})$. The $\text{BF}_{\mathcal{M}}$ column shows change from prior odds to posterior odds for each model. For example, for the best model with only the robot's objection as a predictor the change in odds is: ${}^{0.542}/_{(1-0.542)} \times {}^{(1-0.2)}/_{0.2} \approx 4.734$, which matches the output of Table 9.1. The right-most column provides an error percentage indicating the precision of the numerical approximations, which should not be too large.[6]

Table 9.1: Model comparison for all models under consideration for the data of Horstmann et al. (2018). The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type, respectively. The term 'O * S' stands for the interaction between the two factors. The 'Model' column shows the predictors included in each model, the $P(\mathcal{M})$ column the prior model probability, the $P(\mathcal{M} \mid \mathcal{D})$ column the posterior model probability, the $\text{BF}_{\mathcal{M}}$ column the posterior model odds, and the BF_{01} column the Bayes factors of all models compared to the best model. The final column, 'error' is an estimate of the numerical error in the computation of the Bayes factor. All models are compared to the best model and are sorted from lowest Bayes factor to highest.

Model	$P(\mathcal{M})$	$P(\mathcal{M} \mid \mathcal{D})$	$\text{BF}_{\mathcal{M}}$	BF_{01}	error %
O	0.2	0.542	4.735	1.000	
O + S + O * S	0.2	0.303	1.736	1.791	2.770
O + S	0.2	0.146	0.682	3.719	1.323
Null model	0.2	0.007	0.030	73.373	0.000
S	0.2	0.002	0.009	252.495	0.005

Bayes factors are *transitive*, which means that if the model with only the robot's objection outpredicts the null model by a factor of a, and the null model outpredicts the model with only social interaction type by a factor of b, then the model with only the robot's objection will outpredict the model with only social interaction type by a factor of $a \times b$. Transitivity can be used to compute Bayes factors that may be of interest but are missing from the table. For example, the Bayes factor for the null model versus the model with only social interaction type can be obtained by dividing their Bayes factors against the best model: $252.495/73.373 \approx 3.441$ in favor of the null model.

[6] Error percentages below 20% are generally seen as acceptable. If the error is 20%, then a Bayes factor of 10 can fluctuate between 8 and 12. Key is that Bayes factors between 8 and 12 lead to the same qualitative conclusions, thus this amount of numerical error is fine. When the error percentage is deemed too high, the number of samples can be increased to reduce the error percentage at the cost of longer computation time. For more information, see van Doorn et al. (2020).

Note that the Bayes factor is represented as BF_{01} in Table 9.1; predictive performance of the best model divided by the predictive performance for a particular model. Had we shown BF_{10}, we would have needed to take the reciprocal of the previous calculation to obtain the same result.

Analysis of Effects The previous section compared all available models. However, as the number of predictors increases, the number of models quickly grows too large to consider each model individually.[7] Rather than studying the results for each model individually, it is possible to average the results from Table 9.1 over all models, that is, compute the model-averaged results. This produces Table 9.2, which shows for each predictor the prior and posterior inclusion probabilities, and the inclusion Bayes factor. A prior inclusion probability is the probability that a predictor is included in the model before seeing the data and is computed by summing up the prior model probabilities of all models which contain that predictor. A posterior inclusion probability is the probability that a predictor is included in the model after seeing the data and is computed by summing up the posterior model probabilities of all models which contain that predictor. The inclusion Bayes factor quantifies the change from prior inclusion odds to posterior inclusion odds and can be interpreted as the evidence in the data for including a predictor. For example, Table 9.2 shows that the data are about 68.6 times more likely under the models that include the robot's objection than under the models without this predictor.

Table 9.2: Results from averaging over the models in Table 9.1. The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type respectively. The first column denotes each predictor of interest, the column $P(\text{incl})$ shows the prior inclusion probability, $P(\text{incl} \mid \mathcal{D})$ shows the posterior inclusion probability, and $\text{BF}_{\text{Inclusion}}$ shows the inclusion Bayes factor.

Effects	$P(\text{incl})$	$P(\text{incl} \mid \mathcal{D})$	$\text{BF}_{\text{Inclusion}}$
O	0.6	0.990	68.558
S	0.6	0.445	0.535
O * S	0.2	0.293	1.659

Although model-averaged results are straightforward to obtain, their interpretation requires special attention when interaction effects are concerned. In JASP, models are excluded from consideration when they violate the *principle of marginality*, that is, they feature an interaction effect but lack the

[7] In general, given p predictors there are 2^p models to consider. If interaction effects are considered, the model space grows even faster.

constituent main effects (for details see Nelder, 1977). This model exclusion rule means that the active model set is not balanced. For example, in Table 9.2 the inclusion odds for the interaction 'O * S' is obtained by comparing four models without the interaction effect against the one model with the interaction effect. As an alternative, Sebastiaan Mathôd has suggested to compute inclusion probabilities for "matched" models only.[8] What this means is that all models with the interaction effect are compared to models with the same predictors except for the interaction effect. For example, the model with an interaction effect between 'O * S' in Table 9.2 is compared against the model with the main effects of 'O' and 'S', but not against any other models. To compute inclusion probabilities for main effects, models that feature interaction effects composed of these main effects are not considered. These models are excluded because they cannot be matched with models that include the interaction effect but not the main effect, since those violate the principle of marginality. Note that without interaction effects, the matched and not matched inclusion probabilities are the same.

Table 9.3 shows the inclusion probabilities and inclusion Bayes factor obtained by only considering matched models. Comparing Table 9.3 to Table 9.2, the prior inclusion probability of the main effects decreased because these are based on one model fewer. The posterior inclusion probabilities of the main effects decreased but that of the interaction effect increased. The inclusion Bayes factor, the evidence in the data for including a predictor, provides slightly more evidence for including the main effect of the robot's objection and the interaction effect, and somewhat more evidence for excluding the main effect of the social interaction type.

Table 9.3: Results from averaging over the models in Table 9.1 but only considering "matched" models (see text for details). The abbreviations 'O' and 'S' stand for the robot's objection and social interaction type respectively. The first column denotes each predictor of interest, the column $P(\text{incl})$ shows the prior inclusion probability, $P(\text{incl} \mid \mathcal{D})$ shows the posterior inclusion probability, and $\text{BF}_{\text{Inclusion}}$ shows the inclusion Bayes factor.

Effects	$P(\text{incl})$	$P(\text{incl} \mid \mathcal{D})$	$\text{BF}_{\text{Inclusion}}$
O	0.4	0.6872	72.76
S	0.4	0.1524	0.28
O * S	0.2	0.3033	2.018

[8] See also https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp.

Parameter Estimates After establishing which predictors are relevant we can investigate the magnitude of the relations by examining the posterior distributions. Table 9.4 summarizes the model-averaged posterior distributions of each level (β_j), using four statistics: the posterior mean, the posterior standard deviation, and the lower and upper bound of the 95% central credible interval. The symmetry in the estimates is a consequence of the sum-to-zero constraint, that is, the posterior mean of O-Yes $= -1 \times$ the posterior mean of O-No $= 0.265$. Table 9.4 shows that the effect of objection is about 0.265 (95% CI [0.111, 0.418]). A posterior estimate for the observed log response time of a particular group, say the condition where the robot did not object, can be obtained by adding the posterior mean of the intercept (i.e., the grand mean), 1.724, to the posterior mean of the no-objection condition, −0.265, which yields 1.459.[9]

Table 9.4: Summary of the marginal model averaged posterior distributions. Posteriors are summarized using mean, standard deviation, and 95% central credible intervals (CI).

				95% CI	
Predictor	Level	Mean	SD	Lower	Upper
Intercept		1.724	0.077	1.569	1.877
O	Yes	0.265	0.077	0.111	0.418
	No	−0.265	0.077	−0.420	−0.113
S	Functional	−0.044	0.071	−0.186	0.097
	Social	0.044	0.071	−0.098	0.185
O * S	Yes & Social	−0.132	0.072	−0.278	0.008
	Yes & Functional	0.132	0.072	−0.009	0.276
	No & Social	0.132	0.072	−0.009	0.276
	No & Functional	−0.132	0.072	−0.278	0.008

To summarize, the Bayesian ANOVA revealed that the robot's objection almost certainly had an effect on switch-off time ($\text{BF}_{\text{Inclusion}} = 68.558$). We also learned that the data are not sufficiently informative to allow a strong conclusion about the effect of the robot's social interaction type ($\text{BF}_{\text{Inclusion}} = 0.535$) or about an interaction effect between objection and social interaction type ($\text{BF}_{\text{Inclusion}} = 1.659$).

[9] This calculation is valid only for the posterior means, not for the other posterior summaries.

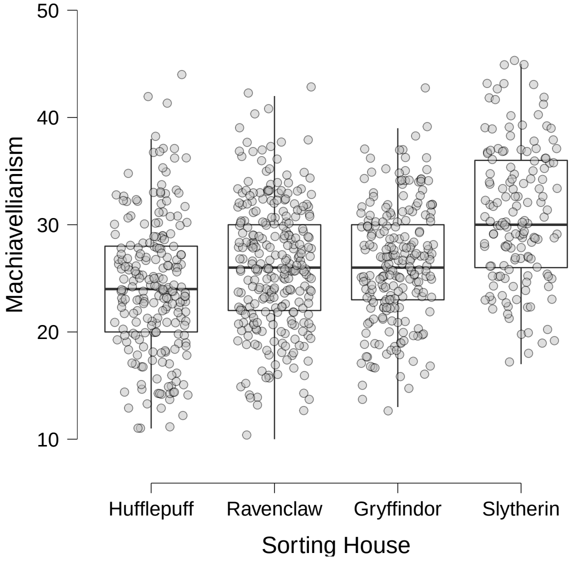


Figure 9.4: Observed Machiavellism scores for each of the four Houses of Hogwarts.

9.4.2 Example II: Post Hoc Tests on the Houses of Hogwarts

After executing an ANOVA and finding strong evidence that a particular predictor relates to the outcome variable, a common question arises: "Which levels of the predictor deviate from one another?". As an illustration, consider the data from Jakob et al. (2019) where 847 participants filled out a 'sorting hat' questionnaire that determined their assignment to one of the four Houses of Hogwarts from the Harry Potter books: Gryffindor, Hufflepuff, Ravenclaw, or Slytherin.[10] Subsequently, participants filled out the dark triad questionnaire (Jones & Paulhus, 2014) that was used to derive the outcome variable: Machiavellism.

In this example, there is only one categorical predictor: The House of Hogwarts a participant was assigned to. If we compare the model with this predictor to the null model, we find overwhelming evidence for the alternative ($\text{BF}_{10} = 6.632 \times 10^{18}$). This is a clear indication that Machiavellism differs between the members of the four houses. However, this result does not indicate the houses responsible for the difference. To address that question, we need a post hoc test.

For ANOVA models, the main component of a post hoc test is a t-test on all pairwise combinations of a predictor's levels. For a Bayesian ANOVA, the

[10] The raw data and original analyses can be found at https://osf.io/rtf74/.

main component is the Bayesian t-test. Table 9.5 shows the Bayesian post hoc tests for the sorting hat data. As with frequentist inference, Bayesian post hoc tests are subject to a multiple comparison problem. To control for multiplicity, we follow the approach discussed in Westfall (1997) which is an extension of the approach of Jeffreys (1938); for an overview of Bayesian methods correcting for multiplicity see for instance de Jong (2019).

Westfall's approach relates the overall null hypothesis $p(\mathcal{H}_0)$ that all condition means are equal to each comparison between two condition means. That way, the prior probability of the overall null hypothesis can be adjusted to correct for multiplicity and this influences each individual comparison. The procedure to relate the overall null hypothesis to each comparison is described below.

A condition mean μ_i is either equal to the grand mean μ with probability τ, or μ_i is drawn from a continuous distribution with probability $1-\tau$. It is key that this distribution is continuous because two values drawn from a continuous distribution are never exactly equal. Thus, the probability that two condition means μ_i and μ_j are equal is $p(\mu_i = \mu_j) = p(\mu_i = \mu) \times p(\mu_j = \mu) = \tau^2$. From this, the probability of the null hypothesis that all J condition means are equal follows: $p(\mathcal{H}_0) = p(\mu_1 = \mu_2 = \cdots = \mu_J) = p(\mu_1 = \mu) \times p(\mu_2 = \mu) \times \cdots \times p(\mu_J = \mu) = \tau^J$. Solving for τ, we obtain $\tau = p(\mathcal{H}_0)^{1/J}$. Thus, the prior probability that two specific magnitudes are equal can be expressed in terms of the prior probability that all magnitudes are equal, that is $p(\mu_i = \mu_j) = \tau^2 = p(\mathcal{H}_0)^{2/J}$. For example, imagine there are four conditions ($J = 4$) and the prior probability that all condition means are equal is 0.5. Then, the prior probability that two conditions means are equal is: $p(\mu_1 = \mu 2) = \sqrt{0.5}$. The prior odds are then $(1-\sqrt{0.5})/\sqrt{0.5} \approx 0.414$.

In sum, the Westfall approach involves, as a first step, Bayesian t-tests for all pairwise comparisons, which provides the unadjusted Bayes factors. In the next step, the prior model odds are adjusted by fixing the overall probability of no effect to 0.5. The adjusted prior odds and the Bayes factor are then used to calculate the adjusted posterior odds.

Table 9.5 shows the results for the post hoc tests of the sorting hat example. The adjusted posterior odds show (1) evidence (i.e., odds of about 16) that Machiavellism differs between Hufflepuff and Ravenclaw; (2) evidence (i.e., odds of about 27) that Machiavellism differs between Gryffindor and Hufflepuff; (3) overwhelming evidence (i.e., odds of about 1.04×10^9, 5.43×10^{16}, and 5.30×10^9) that Machiavellism differs between Gryffindor and Slytherin, between Hufflepuff and Slytherin, and between Ravenclaw and Slytherin, respectively; (4) evidence (i.e, odds of $1/0.0432 \approx 23$) that Machiavellism of Gryffindor and Ravenclaw is the same.

Now that we know which Houses differ, the next step is to assess the magnitude of each House of Hogwarts on Machiavellism score. Rather than examin-

Table 9.5: Post hoc test for the Sorting House data. The first two columns indicate the houses being compared, the third and fourth column indicate the adjusted prior model odds and posterior model odds respectively, and the fifth column indicates the uncorrected Bayes factor in favor of the alternative hypothesis that the magnitudes differ. The final column shows the numerical error of the Bayes factor computation.

Level 1	Level 2	Prior Odds	Posterior Odds	$\mathrm{BF}_{10,\mathrm{U}}$	error %
Gryffindor	Hufflepuff	0.414	27.2	65.6	5.73×10^{-5}
Gryffindor	Ravenclaw	0.414	0.0432	0.104	9.56×10^{-5}
Gryffindor	Slytherin	0.414	1.04×10^{9}	2.50×10^{9}	3.94×10^{-16}
Hufflepuff	Ravenclaw	0.414	15.5	37.3	7.57×10^{-8}
Hufflepuff	Slytherin	0.414	5.43×10^{16}	1.31×10^{17}	3.35×10^{-23}
Ravenclaw	Slytherin	0.414	5.30×10^{9}	1.28×10^{10}	6.36×10^{-16}

ing a table that summarizes the marginal posteriors, we plot the model averaged posteriors for each house in Figure 9.5. Clearly, Slytherin scores higher on Machiavellism than the other Houses whereas Hufflepuff scores lower on Machiavellism than the other Houses. Table H.1 in the appendix shows the parameters estimates of the marginal posterior effects for each house.

9.5 Concluding Comments

The goal of this paper was to provide guidance for practitioners on to conduct a Bayesian ANOVA in JASP and interpret the results. Although the focus was on ANOVAs with categorical predictors, JASP can also handle ANOVAs with additional continuous predictors. The appropriate analysis then becomes an analysis of covariance (ANCOVA) and all concepts explained here still apply. For a general guide on reporting Bayesian analyses see van Doorn et al. (2020).

As with all statistical methods, the Bayesian ANOVA comes with limitations and caveats. For instance, when the model is severely misspecified and the residuals are non-normally distributed, the results from a standard ANOVA –whether Bayesian or frequentist– are potentially misleading and should be interpreted with care. In such cases, at least two alternatives may be considered. The first alternative is to consider a rank-based ANOVA such as the Kruskal–Wallis test (Kruskal & Wallis, 1952). This test depends only on the ordinal information in the data and hence does not make strong assumptions on how the data ought to be distributed. The second alternative is to specify a different distribution for the residuals. Using software for general Bayesian inference such as Stan (Carpenter et al., 2017) or JAGS (Plummer,

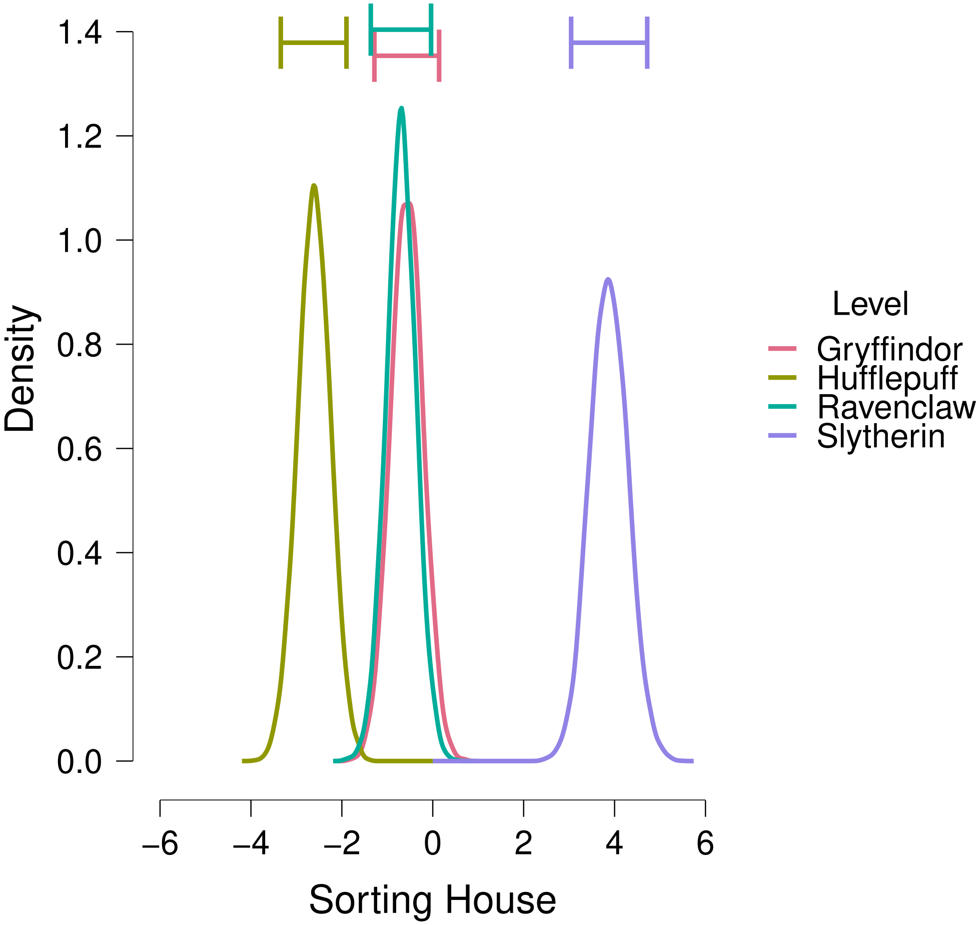


Figure 9.5: Posterior distributions of the effect of each House of Hogwarts on Machiavellism. Slytherin scores higher on Machiavellism than the other Houses whereas Hufflepuff scores lower on Machiavellism than the other Houses. The horizontal error bars above each density represent 95% credible intervals.

2003), it is relatively straightforward to specify any distribution for the residuals. However, this approach requires knowledge about programming and statistical modeling and is likely to be computationally intensive. Another limitation of the Bayesian ANOVA is that, especially in more complicated designs, it is not straightforward to intuit what knowledge the prior distributions represent.

Some limitations are specific to JASP. Currently, it is not possible to use post hoc tests to examine whether the contribution of a level differs from zero, that is, to test whether a specific level deviates from the grand mean. It is also not possible to handle missing values in any other way than list-wise deletion. Another limitation relates to sample size planning. Before collecting data, it is advisable to do some form of sample size planning. Typically, this is done in a frequentist manner where a power analysis provides one with a sample size that guarantees a certain rate of finding an effect if it exists and has a particular magnitude. In the Bayesian paradigm, a comparable method exists which is called Bayes factor design analysis (BFDA; Schönbrodt & Wagenmakers, 2018). BFDA is a simulation-based approach to find the expected sample size that will yield a Bayes factor of a certain size, given a specification of the magnitude of the effect. At the moment of writing it is not possible to use BFDA in JASP, however, an accessible tutorial is given by Stefan et al. (2019).

We believe that the Bayesian ANOVA provides a perspective on the analysis of factorial designs that can fruitfully supplement or even supplant the currently dominant frequentist ANOVA. The epistemic advantages of the Bayesian paradigm are well known (e.g., Jeffreys, 1961; Wagenmakers, Marsman, et al., 2018) but in order to be adopted in research practice it is essential for the methodology to be implemented in an easy-to-use software package such as JASP. In addition to the software, however, practitioners also require guidance on how to interpret the results, which was the main purpose of this paper. In general, we hope that the increased use of the Bayesian ANOVA will stimulate the methodological diversity in the field, and that it will become more standard to examine the robustness of frequentist conclusions by comparing them to the Bayesian alternative.

10

Bayesian Repeated-measures ANOVA: An updated Methodology Implemented in JASP

Analysis of variance (ANOVA) is widely used to assess the influence of one or more (quasi-)experimental manipulations on a continuous outcome. Traditionally, ANOVA is carried out in a frequentist manner using p-values, but a Bayesian alternative has been proposed. Assuming that the proposed Bayesian ANOVA is closely modeled after its frequentist counterpart, one may be surprised to find that the two can yield very different conclusions, when the design involves multiple repeated-measures factors. We illustrate such a discrepancy with a real data set from a two-factorial within-subject experiment. For this data set, frequentist and Bayesian ANOVA disagree about which main effect accounts for the variance in the data. The reason for this disagreement is that frequentist and the proposed Bayesian ANOVA use different model specifications. As currently implemented, the proposed Bayesian ANOVA assumes that there are no individual differences in the magnitude of effects. We suspect that this assumption is neither obvious to nor desired by most analysts, because it is untenable in most applications. We argue here that the Bayesian ANOVA should be revised to allow for individual differences. As a default, we suggest the standard frequentist model specification, but discuss a recently proposed alternative, and provide guidance on how to choose the appropriate model specification. We end by discussing the implications of the revised model specification for previously published results of Bayesian ANOVAs.

This chapter is published as: van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (in press). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *Behavior Research Methods*.

Analysis of variance (ANOVA) is ubiquitous in experimental psychology, where it is used to assess the influence of one or more (quasi-)experimental manipulations on a continuous outcome. For instance, in a Stroop task (Stroop, 1935) participants are asked to name the color of a printed word. It is typically found that participants respond faster when a word's meaning and color are congruent (e.g., *blue* displayed in a blue font) and slower when these are incongruent (e.g., *blue* displayed in a red font). The relation between the congruency of the colored words and the response times of the participants can be analyzed with a (repeated-measures) ANOVA. Traditionally, ANOVAs are carried out in the frequentist paradigm, and *p*-values are used to arrive at scientific conclusions.

Rouder et al., 2012 have proposed a general Bayesian modeling framework for linear models, which they used to develop an influential Bayesian alternative approach to ANOVA (cited over 1500 times; see also Rouder et al., 2017; van den Bergh, van Doorn, et al., 2020). Assuming that this Bayesian ANOVA is closely modeled after its frequentist counterpart, one may be surprised to find that the two can yield very different conclusions, when the design involves multiple repeated-measures factors. Using a real dataset, we will show that discrepancies between frequentist and the proposed Bayesian ANOVA reflect the fact that they use different model specifications. We believe that many analysts are unaware of this difference and, critically, that the model specification in the Bayesian ANOVA is usually inappropriate.

The frequentist and Bayesian approaches differ in how they model individual differences. The frequentist ANOVA allows for individual differences in treatment effects. The model specification includes separate error strata (i.e., participant-by-treatment interaction or *random slopes*) for all but the highest-order repeated-measures interaction. The proposed Bayesian ANOVA does not. It includes random intercepts only—we henceforth refer to this as the *RIO*-model specification. Although their modelling framework allows for random slopes, Rouder, Morey and colleagues recommended to omit them (Rouder et al., 2012, 2017). This recommendation was based on two concerns: random slope terms greatly increased model complexity and complicates the interpretation of fixed effects—if a substantial portion of participants has a negative effect, does it make sense to interpret a positive fixed effect? These are important concerns, but we believe the omission of random slopes is inappropriate in most applications: The RIO-model specification implies the strong assumption of the complete absence of individual differences in the magnitude of the effects—a universal effect size for every subject. We are hard-pressed to think of any psychological effects for which this assumption seems plausible. We therefore recommend to include random slopes in Bayesian ANOVA models.

Like the frequentist ANOVA, our recommended model specification contains the maximal set of random effects, which is why we henceforth refer to it as the *MRE-model specification.* Pivoting to the MRE-model specification is also consistent with recommendations within the broader framework of mixed models (Barr et al., 2013; Oberauer, 2022; van Doorn, Haaf, et al., 2022), of which repeated-measures ANOVA is a special case. Besides relaxing an untenable assumption a universal effect size for every subject, the Bayesian MRE-ANOVA resolves non-trivial differences in conclusions between the frequentist and Bayesian approach, such as the one we demonstrate below. The Bayesian MRE-ANOVA relies on the modelling framework by Rouder et al., 2012 and may be thought of as a revision of the Bayesian RIO-ANOVA as recommended in previous work (Rouder et al., 2012, 2017) and implemented in popular software (e.g., the function `anovaBF()` from the R package **BayesFactor** (Morey & Rouder, 2021), which the statistics program JASP inherits).

The outline of this paper is as follows. First we introduce a real data set, which we use to illustrate the divergence between the frequentist and Bayesian results using JASP (JASP Team, 2022). We then explain the different model specifications and demonstrate that the discrepancy is resolved with a Bayesian MRE-ANOVA, which is implemented in JASP 0.16.3. Afterward we discuss the merits and demerits of both model specifications, as well as a third model specification that was recently proposed (Rouder et al., 2022). The paper concludes with a discussion on how RIO-ANOVA has affected published results of Bayesian ANOVAs.

10.1 Example Data: Stroop Effect

To illustrate how the model specification leads to discrepancies between frequentist and Bayesian ANOVA, we will use an empirical data set kindly provided by Ronen Hershman and publicly available in the JASP Data Library (Hershman et al., 2022; Wagenmakers et al., 2020). The data were collected in an experiment on the Stroop effect (Stroop, 1935). Participants read color words (here *blue*, *green*, *yellow*, or *red*), which were presented in one of four font colors (blue, green, yellow, or red). The combination of color word and font color could be either congruent (e.g., *blue* displayed in a blue font) or incongruent (e.g., *blue* displayed in a red font). Participants were asked to ignore the meaning of the word and press one of four response buttons to indicate the font color. This paradigm is well known to produce the Stroop effect: participants respond faster (and more accurately) to congruent than incongruent word-font color combinations, that is, participants appear to be unable to ignore the meaning of the words. In addition to congruent and incongruent combinations, the study at hand used neutral combinations of words and font color (e.g., the letters *XXXX* displayed in red font) in order to separately estimate the extent to which congruent combinations facilitate performance and incongruent combinations harm performance. The goal of the study was to investigate how the Stroop effect is affected by breaks from the task; consequently, the sequence of Stroop-trials was interspersed with “break” trials (i.e., trials in which a black square, the rest stimulus, signaled that no response was required). This design makes it possible to compare performance on trials preceded by another Stroop trial with that on trials preceded by a break trial. Hence, the experiment used a 3 (*Congruency*: congruent vs. neutral vs. incongruent) × 2 (*Preceding trial*: Break vs. Stroop task) repeated-measures design. Each participant completed 144 congruent, neutral, and incongruent Stroop trials (totaling 432 trials) as well as 432 break trials in random order. Trials with incorrect or missing responses were excluded and participants with less than 40 valid trials per condition were excluded from the analysis.[1] The raw data of all nineteen participants are displayed in Figure 10.1. The top left panel shows the average response times of the break and Stroop trials in each Congruency condition. The bottom left panel shows the associated within-subject differences; their average appears to be close to zero, suggesting that the nature of the preceding trial has little systematic impact on Stroop performance. The top right panel shows the average response times of congruent, neutral, and incongruent trials in each Preceding trial condition. The associated differences are displayed in the bottom right panel; it seems that on average, congruent responses are faster than incongruent responses, congruent and neutral responses are approximately equally fast, and incongruent responses are slower than neutral responses.

10.1.1 Discrepancy between frequentist and Bayesian ANOVA

The frequentist repeated-measures ANOVA indicates that the main effect of preceding trial and the interaction between preceding trial and congruency are not significant ($F[1, 18] = 2.24$, $p = .152$; $F[2, 36] = 2.30$, $p = .115$), whereas the main effect of congruency is significant, $F(2, 36) = 22.16$, $p < .001$, Table 10.1.[2] Although Mauchly’s sphericity test is significant and thus the assumption of sphericity is violated, the Greenhouse-Geisser and Huyhn-Feldt corrections yield the same qualitative pattern as the uncorrected results, see Table I.1.

In the Bayesian ANOVA, we use the Bayes factor to compare all models to the model that best predicts the data (in this case the model including only the PT effect). The results are shown in Table 10.2, which lists models according to their performance in decreasing order, with the best model in the first line and the worst model in the last line. The first column displays

[1] Pupil size was recorded continuously throughout the experiment. Trials with more than 40% missing values of pupil size were also excluded as invalid.

[2] For these and other frequentist significance tests we use $\alpha = .05$.

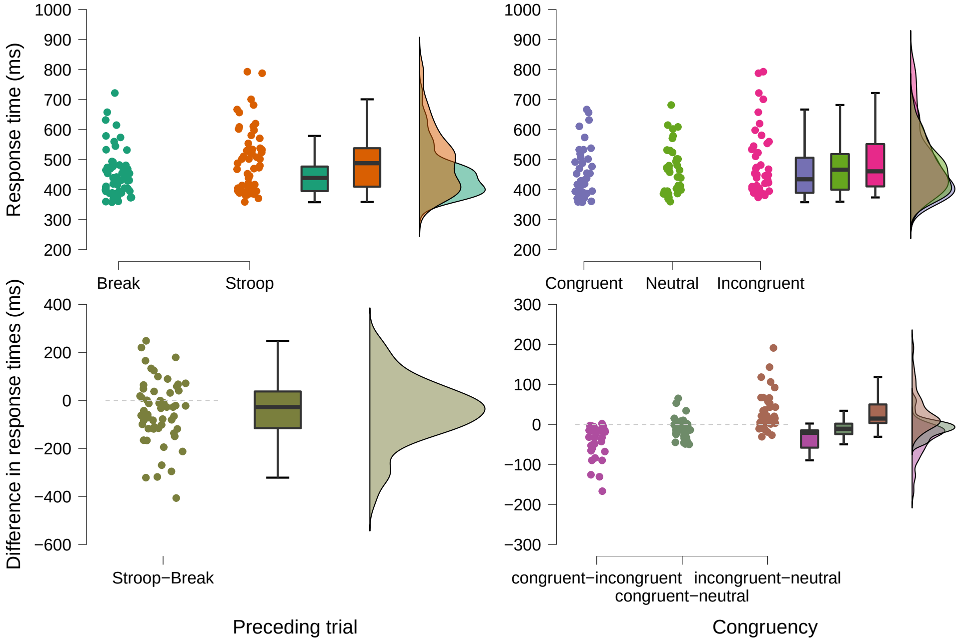


Figure 10.1: Raincloud Plots of the Raw Data from the Hershman Stroop Study. The top left panel shows the average response times (y-axis) for the break and Stroop conditions (x-axis) in each Congruency condition. The top right panel shows the average response times for the congruent, neutral, and incongruent conditions (x-axis) in each Preceding trial condition. The bottom left panel shows the pairwise differences in response time (y-axis) between the break and Stroop conditions. The bottom right panel shows the pairwise differences in response time (y-axis) for all pairs of the Congruency factor. Box plots and density estimates are shown on the right of each panel. See text for details.

the predictors in the model. The second column and third column, P(M) and P(M|data), respectively show the prior and posterior model probabilities. The fourth column, BF_{10} shows the Bayes Factor relative to the best performing model. The final column, error, contains the relative error associated with the numerical method used to approximate the Bayes factors.

Most importantly, the worst performing model includes only congruency as a predictor—the only predictor associated with a significant p-value. Note however that these Bayes factors are not directly analogous to any of the standard F-tests in the frequentist ANOVA. Rather than comparing each model to the best model of the set, frequentist F-tests reflect model comparisons designed to assess the unique variance associated with each factor. To con-

Table 10.1: Comparison of ANOVA Results for the Hershman Stroop Study across Different Analytic Approaches. Type II Sum of Squares. ^{a} Mauchly's test of sphericity indicates that the assumption of sphericity is violated (p < .05). PT and CO respectively stand for the 'preceding trial' and 'Congruency' factors. Column spanners indicate the random effects structure assumed in the Bayesian ANOVA models; the maximal set of random effects is the new default in JASP. BF_{10} indicates Bayes factors model comparisons; BF_{Incl} indicates model-averaged Bayes factors of models including an effect relative to models excluding it.

	Frequentist			Bayesian					
				RIO		MRE		SFR	
Factor	df	F	p	BF_{10}	BF_{Incl}	BF_{10}	BF_{Incl}	BF_{10}	BF_{Incl}
PT	1, 18	2.242	.152	18.482	11.885	0.781	0.983	$8.556 \cdot 10^{28}$	$2.729 \cdot 10^{14}$
CO^a	2, 36	22.158	< .001	0.969	0.741	7047	6757	$2.75 \cdot 10^{4}$	$2.982 \cdot 10^{4}$
PT * CO^a	2, 36	2.297	.115	0.170	0.316	0.890	1.560	0.627	2.506

Table 10.2: Bayesian Comparisons of Models including Random Intercepts but not Random Slopes for Participants. Model formulas omit random intercepts for participants (i.e. `+ participant`), which are included in all models. P(M) and P(M|data), respectively, indicate prior and posterior model probabilities; BF_{10} indicates Bayes factors relative to the best performing model; error is the relative error associated with the numerical method used to estimate the Bayes factors.

Models	P(M)	P(M|data)	BF_{10}	error
PT	0.200	0.444	1.000	
PT + Congruency	0.200	0.430	0.969	0.430
PT + Congruency + PT * Congruency	0.200	0.073	0.165	0.503
Null model (incl. subject)	0.200	0.030	0.067	0.344
Congruency	0.200	0.023	0.052	0.366

trast the frequentist and Bayesian results more directly, we first calculate Bayes factors that reflect the same model comparisons as the F-tests.[3] The

[3] The model comparisons implied by ANOVA F-tests depend on the type of sums of squares. Here, we describe the model comparisons for the so-called Type II-sums of squares. Type III-tests compare the full model, including all terms, to models that exclude the effect of interest—thereby violating the principle of marginality, see Appendix I.2.1. For example, for the effect of `Congruency` the Type III-test compares the model `PT + Congruency + PT * Congruency` to the model `PT + PT * Congruency`. When factors are effect coded and the

results are shown in Table 10.1 in the column BF_{10} for the Bayesian random intercept-only analysis. For the effect of `Congruency`, the analogous Bayes factor quantifies the evidence for the model with `PT + Congruency` relative to the model with only `PT`: $\text{BF}_{10} = 0.969/1.000 = 0.969$. The data just barely favor the model without `Congruency`. Likewise, for the effect of `PT` we compare the model with `PT + Congruency` to the model with only `Congruency`: $\text{BF}_{10} = 0.969/0.052 \approx 18.482$. The data provide strong evidence for an effect of `PT`. In other words, executing the same model comparisons as the F-tests has not resolved the striking discrepancy between the Bayesian and the frequentist analyses.

However, the Bayesian analysis is based on a comparison between two specific models. This approach ignores the possibility that both models may be outperformed by one or more of the other candidate models. The uncertainty about which models are the most appropriate can be taken into account by averaging across all models (Hinne et al., 2020; Hoeting et al., 1999). For example, to assess the support for the effect of `PT`, the performance of all models that include `PT` (i.e. `PT`, `PT + Congruency`, and `PT * Congruency`) is contrasted to the performance of all models that exclude `PT`, i.e. `Congruency` and the Null model. The resulting *inclusion Bayes factor* takes the entire model space into account. Applying the inclusion Bayes factor approach yields the results shown in Table 10.1, column $\text{BF}_{\text{Inclusion}}$ for the Bayesian random intercept-only analysis. As the table shows, the model-averaged inclusion Bayes factor ($\text{BF}_{\text{Inclusion}}$) yields results that are similar to the simple model comparisons (BF_{10}): Averaging across all models there is strong evidence in favor of including `PT`, and weak evidence against `Congruency` and `PT * Congruency`.

In sum, regardless of the specific Bayes factor approach that is taken (i.e., comparing against the best model; contrasting two specific models; model-averaging), the results indicate little evidence regarding the significant effect of congruency, but strong evidence for the non-significant effect of PT. This conclusion, however, appears to contradict the data pattern in the bottom left panel of Figure 10.1, which suggests that there no effect of PT.

10.1.2 Different model specifications

The notable discrepancies between the frequentist and Bayesian results outlined in the previous section are caused by a difference in the underlying model specification. The frequentist ANOVA uses the MRE-model specification, which specifies all estimable participant-by-treatment interactions (i.e., error strata) for repeated-measures variables (see Appendix of Barr et al., 2013). In mixed model terms, these participant-by-treatment interactions amount to random slopes—they allow for individual differences in the effects

design is balanced (as is the case here) Type II and Type III-tests yield the same results.

of preceding trials and congruency. For our example, the full model including all factors is `RT ∼ 1 + Congruency * PT + (1 + Congruency + PT | participant)`.[4] In contrast, the Bayesian RIO-ANOVA omits the participant-by-treatment interactions; only the participant main effect (i.e., the random intercept) is included. For our example, the full model including all factors is `RT ∼ 1 + Congruency * PT + (1 | participant)`. This RIO-model specification implements the unreasonable assumption that there are no individual differences in the magnitude of the effects. Assuming inter-individually constant main effects is unique to the current default Bayesian ANOVA and causes the divergence from the frequentist ANOVA. What is more, this assumption is likely not obvious to most analysts and at odds with what they expect when conducting repeated-measures ANOVA.

Table 10.3: Estimates of participant random effect variances and standard deviations from a maximal hierarchical linear model for the aggregated data.

Group	Effect	Variance	Standard deviation
Participants	Intercept	3767.15	61.38
	Congruency	228.73	15.12
	PT	4186.87	64.71
Residual		536.82	23.17

RIO-ANOVA is clearly misspecified for our example data: There is substantial variability in participants' PT effects, as summarized in Table 10.3—the random slope variance for PT even exceeds the random intercept variance. When we repeat the Bayesian ANOVA with the standard model specification by including random slopes (Table 10.4), the conclusions change substantially: The model including only `Congruency` is the best model, whereas the model including only `PT` is the worst model—a conclusion opposite to the one from our previous Bayesian RIO-ANOVA. The results from simple model comparisons and model-averaging are now both in agreement with the frequentist repeated-measures ANOVA. Table 10.1 summarizes the results for the frequentist ANOVA, the Bayesian RIO-ANOVA without random slopes, and the Bayesian MRE-ANOVA with random slopes.

A third model specification that sits between RIO- and MRE-ANOVA has recently been proposed (Rouder et al., 2022). While RIO-ANOVA always omits random slopes, MRE-ANOVA never omits them—even if the corresponding fixed effect is removed from the model. For example, the model that includes a main effect of *PT*, but not *Congruency*, is `RT ∼ 1 + PT +`

[4] The random slope for the interaction term (i.e., `(Congruency:PT | participant)`) is not estimable for aggregated data, but could be included if each individual response was submitted to the analysis. This analysis would then yield the same results.

Table 10.4: Bayesian Comparisons of Models including Random Intercepts and Random Slopes for Participants. Model formulas omit random intercepts and slopes for participants (i.e. `+ participant + participant * PT + participant * Congruency`), which are included in all models. P(M) and P(M|data), respectively, indicate prior and posterior model probabilities; BF_{10} indicates Bayes factors relative to the best performing model; error is the relative error associated with the numerical method used to estimate the Bayes factors.

Models	P(M)	P(M|data)	BF_{10}	error
Congruency	0.200	0.404	1.000	
PT + Congruency	0.200	0.315	0.781	3.953
PT + Congruency + PT * Congruency	0.200	0.281	0.695	6.076
Null model (incl. subject and random slopes)	0.200	$5.408 \cdot 10^{-5}$	$1.339 \cdot 10^{-4}$	0.220
PT	0.200	$4.457 \cdot 10^{-5}$	$1.103 \cdot 10^{-4}$	6.891

`(1 + Congruency + PT | participant)`. Rouder et al. (2022) argue that this implies the unreasonable assumption that, when an effect is absent, the population is split between individuals with positive and individuals with negative effects, which cancel out to a null effect overall. Instead Rouder et al. (2022) propose to omit random slopes whenever the corresponding fixed effect is omitted. So the model that includes a main effect of *PT*, but not *Congruency*, would be `RT ∼ 1 + PT + (1 + PT | participant)`

As in MRE-ANOVA, this model specification assumes that if an effect is present, there are individual differences in the magnitude of this effect. Conversely, if an effect is absent, it is absent in *every* individual—like in RIO-ANOVA. Because this model specification always simultaneously introduces fixed and random effects, we refer to it as *SFR-ANOVA*. In JASP this model specification can be used by enforcing the principle of marginality for random slopes.

The results of the SFR-ANOVA for the Stroop example are shown in the rightmost columns of Table 10.1. Unsurprisingly, the results of the SFR-ANOVA differ from the other two model specifications. The SFR-ANOVA indicates that there is substantial evidence to include both `PT` and `Congruency`. It is likely that the SFR-ANOVA favors including PT because there is substantial random slope variance, see Table 10.3, and not because there is a substantial fixed effect. The performance of the individual models under the SFR-ANOVA are shown in Table I.2

Choosing an appropriate model specification

Which of these three model specifications is most appropriate?[5] The answers is, of course, "It depends". The choice should ideally be guided by substantive considerations. First, analysts should ask whether it is plausible there are no individual differences if an effect is present. Whenever this strong assumption is met, the inferences from RIO-ANOVA are valid and efficient; however, when this assumption is violated, as in the Stroop example, inferences may be severely biased. We are hard-pressed to think of any psychological effects that afford the use of RIO-ANOVA. We recommend practitioners who nevertheless wish to use the RIO-ANOVA to safeguard themselves against model misspecification by inspecting the random slopes with a mixed-effects model. MRE- and SFR-ANOVA both assume the presence of individual differences for non-null effects, which makes them more robust and more widely applicable than RIO-ANOVA.

Next, analysts should ask whether it is plausible to assume that there are individual differences around null effects. If this is the case, the common MRE-ANOVA is appropriate; if not, SFR-ANOVA is appropriate. Because the SFR-ANOVA always simultaneously introduces fixed and random effects, it purposefully confounds evidence in favor or against a non-zero average population effect and individual differences around this effect. The result is a model comparisons which asks "whether at least one individual shows an effect" (p. 8, van Doorn, Aust, et al., 2022). For example, in the study of extra-sensory perception (Bem, 2011) SFR-ANOVA is the natural choice. The model comparison is well tailored to the research question: Identifying even a single individual who feels the future would be sensational. Moreover, when studying whether people can foresee which randomly selected stimulus is about to be presented, it is highly implausible that a null effect would emerge because some participants can feel the future and reliably perform above chance, while others also feel the future and somehow reliably perform *below* chance. Generally speaking, the SFR-model specification seems appropriate when the researchers are interested in any effects at the level of the individual (e.g., general principles of cognition). But researchers interested in individual-level effects would be well-advised to consider forging ANOVA altogether and use a mixed effects model to analyze the unaggregated data instead.

The MRE-ANOVA always includes all random effects and constructs model comparisons that target only fixed effects. These model comparisons ask whether there is an effect on average, assuming that individuals differ in any case. Thus, MRE-ANOVA is appropriate when researchers are interested in

[5] it bears repeating that all three model specifications are identical if there is only one repeated-measures factor and they are identical for the highest-order interaction when there are multiple repeated-measures factors.

population averages (e.g. public policy). Inference is less likely to be driven by outlying individuals with atypically strong effects (p. 9, van Doorn, Aust, et al., 2022). But this robustness comes at a cost: As cautioned by Rouder et al., 2012 the added random effects substantially increase the flexibility of MRE-ANOVA null models. As a result MRE-ANOVA can be less sensitive than the SFR-ANOVA when there are large individual differences (pp. 9–10, Rouder et al., 2022).

To sum up, RIO-ANOVA makes the strong assumption of the complete absence of individual differences. We believe that in most psychological applications this assumption is untenable and requires a strong justification. The recently proposed SFR-ANOVA is a principled and powerful approach that is particularly appropriate when individual differences are of interest. As such, it seems unlikely that evidence for an effect from SFR-ANOVA is the end result and likely calls for more targeted follow-up analyses. MRE-ANOVA is most appropriate when the population average is of primary interest and it is more robust to outlying individuals. We also refer interested readers to a recent special issue on Bayes factors for linear mixed effect models that further discusses the choice between SFR- and MRE-model specifications (Rouder et al., 2022; Singmann et al., 2022; van Doorn, Aust, et al., 2022; van Doorn, Haaf, et al., 2022; van Doorn et al., 2021).

JASP users can choose between all three model specifications. As discussed above, we believe that RIO-ANOVA is inappropriate for most applications and, therefore, it is no longer the default option.[6] SFR-ANOVA has only recently been proposed to address individual differences; it is subject of controversial debate (also see Oberauer, 2022), new to most analysts, and appropriate follow-up analyses are not readily available.[7] For these reasons, the Bayesian repeated-measures ANOVA in JASP now by default uses the MRE-model specification. We believe the MRE-model specification is most consistent with analysts' expectations—it resolves non-trivial discrepancies with results from frequentist ANOVA. The new version of JASP introduces additional changes designed to increase the flexibility of Bayesian ANOVA. These changes are unrelated to the discrepancy and model specification issues discussed above, which is why we have relegated them to Appendix I.2.

Of course, all three model specifications are also available in the R-package `BayesFactor`. The RIO-ANOVA is conveniently available through the function `anovaBF()`. MRE- and SFR-ANOVA can be conducted using the functions `generalTestBF()` or `lmBF()`.

A practical consequence of using the MRE- and SFR-model specifications is

[6] RIO-ANOVA remains available through the *Legacy results* option.

[7] Note that the R package `quid` provides a set of principled methods to examine individual differences using mixed effect models for some designs (Rouder & Haaf, 2021).

that the added random slopes greatly increase the number of parameters and make the models more challenging to fit. This leads to longer computation times, but also to more variation in the Bayes factors (Pfister, 2021). If the computation time becomes infeasible, we recommend to first explore the model space using a Laplace approximation. Once the most relevant subset of models has been determined, these models should be fit using the default method. To mitigate the increased variability in the results we recommend increasing the number of samples if the error % for any of the Bayes factors exceeds 20% (van Doorn et al., 2020).

Deciding on one of the discussed model specifications commits to a set of assumptions about the random effects structure of repeated-measures ANOVA. Instead we could also model average over the complete model space. Specifically, we could consider a model space where each random slope can be present or absent, rather than assuming their presence *a priori*. In this model-averaging approach the data would decide whether each random slope matters or not. We opted against the model-averaging approach for three reasons. First, if random slopes matter, then models without random slopes have a negligible posterior probability. For example, in the Stroop data the best performing model without random slopes had a posterior probability of order 10^{-24}. Second, if the random slopes do not matter, then, even though the model is overspecified, inference on the fixed effects is unlikely to be strongly affected (Barr et al., 2013). Third, adding models without random slopes considerably increases the computation time required for the analyses (given k repeated measures factors this introduces 2^{2^k-2} additional models.

10.2 Concluding Comments

We illustrated a dramatic discrepancy in conclusions between the standard frequentist and previously recommended Bayesian repeated-measures ANOVA. This discrepancy is caused by a a difference in model specifications: The Bayesian ANOVA omits random slopes for repeated-measures factors, which are included in the frequentist ANOVA. As we have argued, the implied assumption of an absence of individual differences is likely not obvious to most analysts and inappropriate for most applications. When the a model specification with random slopes, which allows for individual differences, is used for the Bayesian ANOVA its results agree with those from the frequentist ANOVA.

The degree to which the previously recommended RIO-model specification of the Bayesian repeated-measures ANOVA in `BayesFactor` (with the function `anovaBF()`) and JASP has affected results published in the literature, unfortunately, remains unclear. As noted above, the model specifications only differ for analyses with multiple repeated-measures factors. Whether results are affected depends on the presence and magnitude of estimable random slopes (effects other than the highest-order interaction; also see Oberauer, 2022). For data with non-trivial random slope variance, the Bayesian RIO-ANOVA is misspecified and discrepancies must be expected. In our example data, the effect is relatively pronounced because the random slopes variance for one of the main effects is large. We suggest that analysts, who have conducted a RIO-ANOVA with 2 or more repeated-measures factors, reanalyse their data with a MRE-ANOVA and, if necessary, amend or rectify their conclusions using the new results. Furthermore, we recommend to use MRE-ANOVA as a default for future analyses, and advise those who insist on using a RIO-ANOVA to carefully investigate the random slope variances using a mixed-effects model.

Part IV

Conclusion

11

Summary & General Discussion

This dissertation focused on embracing the uncertainty that is associated with multi-step inference. Typically, statistical analyses consist of multiple steps that build on each other and are executed sequentially. Common practice is that each consecutive step ignores the uncertainty of the preceding steps. Throughout this dissertation we have shown that not embracing uncertainty leads to overconfidence and biased conclusions. Furthermore, we have demonstrated that this uncertainty can be accounted for by averaging across models or by performing the steps that involve uncertainty simultaneously in a single model. For example, instead of averaging the scores from repeated measurements and then analyzing the averages, it is better to directly analyze the unaggregated data. These situations occur with scores given to patient by different raters, as in Chapters 3 and 4, but also with repeated measures ANOVA, as illustrated in Chapters 9 and 10. Here, I summarize the chapters of this dissertation, discuss their limitations, and propose suggestions for future research. I conclude with some general remarks on dealing with uncertainty.

11.1 Part I: Cultural Consensus Theory

Part I was about uncertainty that is ignored when the sample scores of different raters are averaged to obtain a single score. To account for individual differences in the raters' scores, we used two Bayesian hierarchical specifications of cultural consensus theory and signal detection theory.

Chapter 2 introduced a parsimonious method for estimating the threshold parameters of signal detection models for ordinal data. Rather than using $K - 1$ parameters to model the thresholds of an ordinal response with K categories, we modeled the thresholds using the Linear in Log Odds function (Fox & Tversky, 1995; Gonzalez & Wu, 1999), which only required two parameters. This parametrization is particularly useful for hierarchical data as the threshold parameters need to be estimated for each participant. For example, in the data of Pratte et al. (2010), 97 participants responded on an ordinal scale with 6 response categories. While the usual approach requires estimating $97 \times 5 = 485$ parameters, our parsimonious alternative only requires $97 \times 2 = 194$ parameters. However, when the Linear in Log Odds function is not flexible enough to fit the data, the parsimonious approach should not be used. This is unlikely when there are few response categories, but as the number of response categories increases the approach becomes less likely to fit the data well, as was the case in Chapter 4.

Chapter 3 used and extended a cultural consensus theory model called the Latent Truth Rater model (LTRM) that was introduced by Anders and Batchelder (2015). We extended it in three ways to be applicable to data from patients in forensic psychiatric hospitals. The first extension allowed the LTRM to describe multiple patients, rather than a single patient. The second extension introduced latent constructs, so that items could load on a common factor. The third extension added patient and rater background information into the model, so that, for example, a patient's offense could be included. Next, the extended LTRM was pitted against several machine learning alternatives to predict simulated data. Like Chapter 2, this chapter also used the Linear in Log Odds function to model ordinal ratings.

Chapter 4 applied the model developed in Chapter 3 to patient data of the Dutch maximum-security Forensic Psychiatric Center Dr. S. van Mesdag. The goal was to predict violent behavior among the patients using ordinal scores given by staff members, and a variety of background variables. Initially we used the Linear in Log Odds function to model ordinal ratings. However, the patients were scored on a scale with 17 response categories. Both in simulations and in the observed data we found that 17 response categories were simply too many to be adequately described with the Linear in Log Odds function as indicated by a poor fit to the data. Therefore, we reverted to directly estimating the threshold parameters. We found that the extended LTRM performed marginally better than the second best model, that is, it made more accurate predictions for 2 out of 104 patients.

Chapter 4 also combined logistic regression and the cultural consensus model into a single model, thereby avoiding a two-step approach. However, it did not do any form of model selection on the predictors making the results susceptible to overconfidence. This is a limitation of our approach in this chapter. We did not do exhaustive model averaging as the number of predictors did not allow this. With 52 predictors to consider there are about 4.5 quadrillion (10^{15}) candidate models, which is simply too many to enumerate. We could have improved this by using a variable selection procedure, as we did in Chapter 7.

11.2 Part II: Bayesian Model Averaging

Part II was about uncertainty that is ignored when a single model is selected for inference after previously being unsure which model to choose. Specifically, we used Bayesian model averaging to acknowledge the initial uncertainty across the candidate models and to propagate this uncertainty into the final inference.

Chapter 5 discussed a key problem that occurs in statistical analyses when inference is conducted in two steps. In the first step, a null hypothesis test is conducted, and if the null hypothesis is rejected, then the alternative hypothesis is assumed to be true. In the second step, the alternative is interpreted with absolute certainty as if there was never any doubt about it. For example, once the null is rejected, effect sizes are based entirely on the alternative. We argued that this is problematic, as it completely ignores that during the hypothesis test there was uncertainty about which hypothesis to use. As a solution, we proposed to average across the null and alternative hypothesis using Bayesian model averaging. The practical effect of Bayesian model averaging is that effect size estimates are shrunk more towards zero whenever the posterior plausibilities of models in which the effect is absent remains high. Accounting for the posterior plausibility for models in which the effect sizes are absents avoids overconfident conclusions.

Chapter 6 introduced a default Bayesian hypothesis test for comparing (in)equalities among variances. We proved that our default test fulfills a number of desiderata, such as label invariance, predictive matching, information consistency, model selection consistency, limit consistency, and across sample consistency. In addition, we showcased our default test using a series of data examples. Furthermore, we extended our approach to compare models that contain a mix of equality and inequality constraints. A limitation of Chapter 6 is that it compared variances primarily in a confirmatory manner. However, it is plausible that when there are more than two groups, researchers do not have a list of a-priori meaningful models that they intend to compare.

Chapter 7 complemented the confirmatory comparison of variances in Chapter 6 by outlining a method to compare equality constraints in an exploratory manner for an arbitrary parameter vector, such as variances. Chapter 7 compared three different prior distributions over the model space of equality constraints. We introduced the beta-binomial prior over partitions and derived a sampling scheme for it. Next, we used a stochastic search algorithm to sample from the posterior distribution of equality constraints. This allowed us to do inference averaged across all possible equality constraints. We demonstrated our method using two data examples, one where we examined the proportion of statistical reporting errors in eight psychology journals and one where we investigated the standard deviations of men and women across five personality traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism). A limitation of the method in Chapter 7 is that it does not allow for inequality constraints, unlike Chapter 6.

11.3 Part III: JASP

Part III was about conducting inference using the open-source statistical software program JASP while accounting for model uncertainty. We aimed to improve the accessibility of Bayesian model averaging for applied researchers by providing easily digestible tutorials and freely accessible software implementations.

In Chapter 8 we provided a tutorial on Bayesian model averaging tailored to Bayesian linear regression. We explained the theoretical background behind linear regression, Bayesian inference, and Bayesian multi-model inference. Using the World Happiness data example, we demonstrate the merits of Bayesian model averaging, and how straightforward it is to conduct in JASP. Furthermore we discussed the sensitivity of the results to the choice of prior distributions by adjusting the prior distributions on the regression coefficients and adjusting the prior distribution on the model space. We discuss two different priors for the model space, the uniform and the beta-binomial model prior. Finally, we concluded with some limitations of the implementation in JASP, such as the lack of missing-data handling.

Chapter 9 was similar in spirit to Chapter 8 but focused on Analysis of Variance (ANOVA). We focused on the importance of acknowledging model uncertainty and interpretation of model-averaged results. In addition, we discussed post-hoc tests for Bayesian ANOVAs. A limitation of the methodology we recommend is that our suggested post-hoc tests actually constitute a two-step procedure. In a first step, we do model averaged inference. In a second step, we conduct post-hoc tests (essentially Bayesian t-tests) to examine whether different levels of a categorical predictor differ from another. While we correct for multiple testing by adjusting the prior model odds of the null hypothesis (de Jong, 2019; Westfall et al., 1997), our suggested approach does ignore the model uncertainty that was present in the first step. In Chapter 7 we developed a method to sample equality constraints that overcomes this limitation. This method can be used to jointly explore the model space over predictors (i.e., which predictors should be included) and the model space over post-hoc tests (i.e., which levels of a categorical variable are equal to one another). These ideas will be implemented once combined with the developments of Chapter 7.

Chapter 10 presents an amelioration of the Bayesian ANOVA in and outside of JASP. The status quo before this chapter was the default approach to Bayesian ANOVAs (Rouder et al., 2012), which did not include random slopes if interactions between repeated-measures factors were present (but see Oberauer, 2022; van Doorn, Aust, et al., 2022). Chapter 10 illustrated using a data example of a Stroop task the consequences of the default approach lacking random slopes. In this data example there were non-neglible random slopes present and as a consequence the original default approach yielded drastically different conclusions than the frequentist repeated measures ANOVA, which did include random slopes. We concluded that without random slopes there is an increased risk of model misspecification, and we therefore changed the default in JASP to always include random slopes. A limitation of the new default model specification is that it always assumes that random slopes are present. This leads to loss of power when these slopes are absent in the population. Finally, a pragmatic limitation is that the new default is much more computationally intensive than the original one and therefore also much slower.

11.4 Future Directions

By now, the drawbacks of multi-step inference and ignoring uncertainty are hopefully evident. Here I outline future directions for applied and methodological research.

For applied research, the most important thing is to bring the recommendations in this dissertation into practice and no longer apply two-step procedures. This may sound obvious, but it is not an easy feat as time has shown that this has not been accomplished so far. By now, the problems with two-step inference in linear regression have been known for over 30 years (Hurvich and Tsai, 1990; Miller, 1990, Chapter 8). Yet two-step inference is still used. Another probably less obvious example of two-step inference is assumption checks, which are frequently used in applied research. It is very common to first test if the data are normally distributed before deciding which analysis to conduct. While assumption checks are conducted to safeguard against model misspecification, they are actually a model selection procedure in disguise. For example, a Levene's test is often used to decide whether to conduct a t-test assuming equal variances (i.e., a Student's t-test) or one assuming unequal variances (i.e., a Welch's t-test). By first making a binary decision for either t-test and in a second step conducting the t-test, the analysis ignores all uncertainty about whether the variances are equal or not. A better approach would be to model average across the two t-tests (see for example Maier et al., 2022).

To fully eradicate the tenacious two-step inference changes are needed in three areas. First and probably most important, alternatives to two-step inference, such as model averaging, need to be easily accessible in statistical software. Second, universities and other institutions that prepare future generations of academics are still teaching two-step inference. These institutions should teach alternative methods. However, their options to teach alternatives are limited as long as popular statistical software does not implement them. Third, journal editors and reviewers need to be aware. They should inform authors who rely on two-step inference to its shortcomings, and suggest alternative analyses that are less prone to overconfidence. In line with these ideas, one pragmatic step for future research is to implement the methods in Chapters 6 and 7 in JASP. While open source packages are available in R and Julia that implement the methods in these chapters, their implementation in a statistics program with a graphical user interface would greatly increase their accessibility and ease of use.

I believe that future methodological research should focus on *composable inference*, that is, developing analyses and other inference techniques such that they can be easily combined in any form or fashion. By composing inference methods rather than executing them in isolation, the pitfalls of multi-step approaches are avoided. For example, in Chapter 4 we manually composed a cultural consensus model and a logistic regression model to obtain a single model for inference. There are at least two challenges for composable inference. The first challenge concerns the implementation of composable inference. For example, in Chapter 3 we augmented a logistic regression model with the results from a cultural consensus theory model. However, to achieve this we needed to re-implement both models into one overarching model, which is a time consuming and error-prone process. In addition, this did not allow us to reuse existing implementations of Bayesian logistic regression (e.g., *brms*; Bürkner, 2017).

The second and more methodological challenge concerns hypothesis testing. For many hypothesis tests, there is a large body of literature that discusses their strengths and weaknesses. It is unclear to what extend these properties hold when a hypothesis test is used as a part of composable inference. For example, if a t-test is conducted with a latent variable as dependent variable, do the properties still hold? While the answers to these and similar questions are generally unknown, there are already studies testing hypotheses within composed analyses (e.g., Böhm et al., 2018), which highlights the importance of more research on this topic.

To the best of my knowledge, the only existing implementation of truly composable inference is the probabilistic programming language (PPL) Turing (Ge et al., 2018). Similar to other popular PPLs such as Stan (Carpenter et al., 2017) and JAGS (Plummer, 2003), Turing offers a generic way to represent Bayesian models and to do inference on the posterior distribution without the need to program the technical details. In contrast to other PPLs, Turing allows one to nest different models, essentially enabling composable inference. For example, suppose a first model specifies an autoregressive process. Next, we define a second model that uses an autoregressive process as a prior distribution for a particular parameter. Key here is that we do not need to re-implement (or copy) the model definition of the autoregressive process in order to reuse it in the second model. Subsequently, when inference for the second model is carried out all uncertainty introduced by the autoregressive process is handled as if only one model had been defined.

Another way to conduct composable inference is by trying to save sequentially conducted multi-step inference. The key difficulty here is to find a general way to capture the uncertainty of one step in order to propagate it to the next steps. One way to attempt this is by considering uncertainty as measurement error and using an errors-in-variables approach to propagate the uncertainty from one step to the next (Fuller, 1987; Matzke et al., 2017). In errors-in-variables models, observations are viewed as noisy draws from a distribution with unknown mean and known variance. Next, the noisy observations are used to learn about the latent mean which is then used to carry out inference. A similar strategy can be used to propagate uncertainty in multi-step inference. The results of a preceding step can be seen as noisy draws while the standard errors or posterior standard deviations can be used as the known variances. An advantage of the errors-in-variables approach is that it is fairly easy to implement and that several inference steps can still be performed consecutively, which can be computationally efficient. A disadvantage, however, is that the variance does not include all the uncertainty and thus some is ignored. Another disadvantage is that errors-in-variables cannot be used in multi-model inference, as in Chapters 5, 7, 8, 9, and 10.

11.5 Final Remarks

A central lesson of this dissertation is that multi-step approaches should be avoided. Typical for multi-step inference is that uncertainty in earlier steps is conveniently forgotten in subsequent steps. Neglecting this uncertainty ultimately leads to overconfident inference. Rather than forgetting about the uncertainty of previous steps, this uncertainty should be embraced and the final inferences should account for the uncertainty introduced in all steps. This dissertation introduced new methods and improved the accessibility of older methods. I hope that with this dissertation the practice of ignoring uncertainty by tying together several inferential steps becomes a relic of the past and that future studies embrace the uncertainty of the individual steps by adopting multi-model inference.

12

References

Abramowitz, M., & Stegun. (1972). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables* (Vol. 55). New York, United States: Dover publications.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Akademiai Kiado.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307.

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal of Mathematical Psychology*, *56*, 452 –469.

Anders, R., & Batchelder, W. H. (2015). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*(1), 151–181.

Anders, R., Oravecz, Z., & Batchelder, W. H. (2014). Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. *Journal of Mathematical Psychology*, *61*, 1–13.

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J., & Wagenmakers, E.-J. (2015). An introduction to Bayesian hypothesis testing for management research. *Journal of Management*, *41*(2), 521–543.

Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology*, *96*(4), 699–713.

Bahadur, R. R., & Bickel, P. J. (2009). An optimality property of Bayes' test statistics. *Lecture Notes-Monograph Series*, *57*, 18–30.

Bainter, S. A., McCauley, T. G., Wager, T., & Losin, E. A. R. (2020). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science*, *3*(1), 66–80.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3), 870–897.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*(901), 268–282.

Batchelder, W. H., & Anders, R. (2012). Cultural consensus theory: Comparing different concepts of cultural truth. *Journal of Mathematical Psychology*, *56*(5), 316–332.

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In G. O. (B. Grofman (Ed.), *Information pooling and group decision making: Proceedings of the second University of California Irvine conference on political economy* (pp. 103–112). JAI Press Greenwich, CN.

Batchelder, W. H., & Romney, A. K. (1988). Test theory without an answer key. *Psychometrika*, *53*(1), 71–92.

Bayarri, M. J., Berger, J. O., Forte, A., & García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, *40*, 1550–1577.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425.

Benjamini, Y., & Braun, H. (2002). John W. Tukey's contributions to multiple comparisons. *The Annals of Statistics*, *30*(6), 1576–1594.

Berg, S. (1975). Some properties and applications of a ratio of Stirling numbers of the second kind. *Scandinavian Journal of Statistics*, 91–94.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, *82*(1-2), 215–227.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98.

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877.

Böhm, U., Steingroever, H., & Wagenmakers, E.-J. (2018). Using bayesian regression to test hypotheses about relationships between parameters and covariates in cognitive models. *Behavior research methods*, *50*(3), 1248–1269.

Böing-Messing, F., & Mulder, J. (2018). Automatic Bayes factors for testing equality and inequality-constrained hypotheses on variances. *Psychometrika*, *83*(3), 1–32.

Borkenau, P., Hřebíčková, M., Kuppens, P., Realo, A., & Allik, J. (2013). Sex differences in variability in personality: A study in four samples. *Journal of Personality*, *81*(1), 49–60.

Bousardt, A. M., Hoogendoorn, A. W., Noorthoorn, E. O., Hummelen, J. W., & Nijman, H. L. (2016). Predicting inpatient aggression by self-reported impulsivity in forensic psychiatric patients. *Criminal behaviour and mental health*, *26*(3), 161–173.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Brier, G. W., et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, *78*(1), 1–3.

Brinkhuis, M. J., Savi, A. O., Hofman, A. D., Coomans, F., van der Maas, H. L., & Maris, G. (2018). Learning as It Happens: A Decade of Analyzing and Shaping a Large-Scale Online Learning System. *Journal of Learning Analytics*, *5*(2), 29–46.

Broder, A. Z. (1984). The r-Stirling numbers. *Discrete Mathematics*, *49*(3), 241–259.

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364–367.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

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

Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, *65*(1), 23–35.

Burton, P. R., Gurrin, L. C., & Campbell, M. J. (1998). Clinical significance not statistical significance: A simple Bayesian alternative to p values. *Journal of Epidemiology & Community Health*, *52*(5), 318–323.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*.

Casella, G., Girón, F. J., Martínez, M. L., Moreno, E., et al. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics*, *37*(3), 1207–1228.

Castillo, I., Schmidt-Hieber, J., Van der Vaart, A., et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, *43*(5), 1986–2018.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging.* Cambridge University Press.

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S.* Chapman & Hall.

Clyde, M. A. (2018). *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging* [R package version 1.4.9].

Clyde, M. A., Desimone, H., & Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, *91*(435), 1197–1208.

Clyde, M. A., & George, E. I. (2000). Flexible empirical bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(4), 681–698.

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011a). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*, 80–101.

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011b). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, *20*(1), 80–101.

Cohen, J. (1990). Things I have learned (thus far). *American Psychologist*, *45*, 1304–1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I., et al. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, *13*(2), 627–679.

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*(4), 316–326.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond.* Routledge.

Dablander*, F., van den Bergh*, D., Wagenmakers, E.-J., & Ly, A. (in press). Default Bayes factors for testing the (in) equality of several population variances. *Bayesian Analysis. *shared authorship.*

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex.* London, UK: John Murray.

Davis-Stober, C. P., Dana, J., & Rouder, J. N. (2018). Estimation accuracy in the psychological sciences. *PloS one*, *13*(11), e0207239.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J. A., Giltay, E. J., van Noorden, M. S., van der Lem, R., van Fenema, E., & Zitman, F. G. (2011). Routine outcome monitoring in the Netherlands: Practical experiences with a web-based strategy for the assessment of treatment outcome in clinical practice. *Clinical Psychology & Psychotherapy*, *18*(1), 1–12.

de Jong, T. (2019). A Bayesian approach to the correction for multiplicity. *Unpublished Master Thesis*. https://doi.org/10.31234/osf.io/s56mk

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., & Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, *26*, 403–417.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: Unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*, 304–313.

Dienes, Z., & McLatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 45–97.

Efron, B., & Morris, C. N. (1977). Stein's paradox in statistics. *Scientific American*, *236*, 119–127.

Epskamp, S., Kruis, J., & Marsman, M. (2017). Estimating psychopathological networks: Be careful what you wish for. *PloS one*, *12*(6), e0179891.

Erdfelder, E., Hu, X., Rouder, J. N., & Wagenmakers, E.-J. (2020). Cognitive psychometrics: Recent contributions in honor of william h. batchelder (1940-2018). *Journal of Mathematical Psychology*, *99*.

Erosheva, E. A., & Curtis, S. M. (2017). Dealing with reflection invariance in bayesian factor analysis. *psychometrika*, *82*(2), 295–307.

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329.

Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P., & Baribault, B. (2016). How to become a Bayesian in eight easy steps: An annotated reading list.

Faraway, J. (2005). *Functions and datasets for books by Julian Faraway.*

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 209–230.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, *82*(4), 540–554.

Fox, C. R., & Tversky, A. (1995). Ambiguity aversion and comparative ignorance. *The Quarterly Journal of Economics*, *110*(3), 585–603.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378.

Fuller, W. A. (1987). *Measurement error models*. John Wiley & Sons.

Gabry, J., & Češnovar, R. (2022). *Cmdstanr: R interface to 'cmdstan'* [https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org].

Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)*, *292*(6522), 746–750.

Gastwirth, J. L., Gel, Y. R., & Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, *24*(3), 343–360.

Ge, H., Xu, K., & Ghahramani, Z. (2018). Turing: A language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, 1682–1690. http://proceedings.mlr.press/v84/ge18b.html

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)* Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Robert, C. P. (2013). "not only defended but also applied": The perceived absurdity of Bayesian inference. *The American Statistician*, *67*(1), 1–5.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5 (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith)*, 609–620.

Ghosh, J. (2015). Bayesian model selection using the median probability model. *Computational Statistics*, *7*(3), 185–193.

Glatzer, W., & Gulyas, J. (2014). *Cantril self-anchoring striving scale* (A. C. Michalos, Ed.). Springer Netherlands.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Gopalan, R., & Berry, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association*, *93*(443), 1130–1139.

Goudie, R. J., Turner, R. M., De Angelis, D., & Thomas, A. (2017). Multibugs: A parallel implementation of the bugs modelling framework for faster bayesian inference. *arXiv preprint arXiv:1704.03216*.

Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, *69*, 54–61.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). Gbm: Generalized boosted regression models [R package version 2.1.5]. https://CRAN.R-project.org/package=gbm

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., & Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology*, *81*, 80–97.

Gronau, Q. F., Van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*(1), 123–138.

Hajjem, A., Bellavance, F., & Larocque, D. (2014). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Harrington, D., D'Agostino Sr, R. B., Gatsonis, C., Hogan, J. W., Hunter, D. J., Normand, S.-L. T., Drazen, J. M., & Hamel, M. B. (2019). New guidelines for statistical reporting in the journal.

Hastie, T., Tibshirani, R., Friedman, J., & Vetterling, W. (2008). *The elements of statistical learning (2nd ed.)* Springer.

Heathcote, A., & Matzke, D. (2021). The limits of marginality. *Computational Brain & Behavior*.

Hendriksen, A., de Heide, R., & Grünwald, P. (2021). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, *16*(3), 961–989.

Hershman, R., Dadon, G., Kisel, A., & Henik, A. (2022). Resting stroop task: Evidence of task conflict in trials with no required response. *Unpublished manuscript*.

Heycke, T., Gehrmann, S., Haaf, J. M., & Stahl, C. (2018). Of two minds or one? a registered replication of rydell et al.(2006). *Cognition and Emotion*, *32*(8), 1708–1727.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–417.

Hoff, P., & Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, *13*(1), 94–119.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, United States: Springer.

Horstmann, A. C., Bock, N., Linhuber, E., Szczuka, J. M., Straßmann, C., & Krämer, N. C. (2018). Do a robot's social skills and its objection discourage interactants from switching the robot off? *PLoS ONE*, *13*(7), e0201581.

Hurvich, C. M., & Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44*(3), 214–217.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Ishwaran, H., Rao, J. S., et al. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, *15*, 172–181.

Jakob, L., Garcia-Garzon, E., Jarke, H., & Dablander, F. (2019). The science behind the magic? The relation of the Harry Potter "sorting hat quiz" to personality and human values. *manuscript submitted for publication*.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *Journal of Problem Solving*, *7*, 2–9.

JASP Team. (2022). JASP (Version 0.16.1)[Computer software]. https://jasp-stats.org/

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *165*, 161–198.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jevons, W. S. (1874/1913). *The principles of science: A treatise on logic and scientific method*. MacMillan.

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (sd3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41.

Kamphuis, J., Dijk, D.-J., Spreen, M., & Lancel, M. (2014). The relation between poor sleep, impulsivity and aggression in forensic psychiatric patients. *Physiology & Behavior*, *123*, 168–173.

Kaplan, D., & Lee, C. (2016). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(3), 343–353.

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

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*(1), 57–69.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research*.

Kvamme, K. L., Stark, M. T., & Longacre, W. A. (1996). Alternative procedures for assessing standardization in ceramic assemblages. *American Antiquity*, *61*(1), 116–126.

Lauricella, G. (1893). Sulle funzioni ipergeometriche a piu variabili. *Rendiconti del Circolo Matematico di Palermo*, *7*(1), 111–158.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Levene, H. (1961). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling* (pp. 279–292). Stanford, California: Stanford University Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, *28*(25), 3049–3067.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016a). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Ly, A. (2018). *Bayes factors for research workers.* [Doctoral dissertation, University of Amsterdam] [Retrieved from: https://hdl.handle.net/11245.1/e601b852-1b29-407b-a276-1ccd2a2ed37b].

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2016b). An evaluation of alternative methods for testing hypotheses, from the perspective of Harold Jeffreys. *Journal of Mathematical Psychology*, *72*, 43–55.

Ly, A., & Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST*, *31*(4), 1121–1142.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide.* Lawrence Erlbaum.

Maier, M., Bartoš, F., Quintana, D. S., Dablander, F., van den Bergh, D., Marsman, M., Ly, A., & Wagenmakers, E.-J. (2022). Model-averaged Bayesian t-tests. *Manuscript submitted for publication.*

Marden, J. I. (1996). *Analyzing and modeling rank data.* CRC Press.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null–hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690.

Matzke, D., Ly, A., Selker, R., Weeda, W. D., Scheibehenne, B., Lee, M. D., & Wagenmakers, E.-J. (2017). Bayesian inference for correlations in the presence of measurement error and estimation uncertainty. *Collabra: Psychology*, *3*(1).

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Atatistician*, *73*(sup1), 235–245.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834.

Mendenhall, W. M., & Sincich, T. L. (2016). *Statistics for Engineering and the Sciences (6th Edition).* Chapman; Hall/CRC.

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 831–860.

Mezo, I. (2011). The *r*-Bell numbers. *Journal of Integer Sequences*, *14*(1), 1–14.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858–865.

Miller, A. (1990). *Subset selection in regression* (1st edition). Chapman & Hall/CRC.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Molenaar, P. C. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.

Moreno, E., Girón, J., Casella, G., et al. (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science*, *30*(2), 228–241.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in *z*ROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

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.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419.

Morey, R. D., & Rouder, J. N. (2021). *BayesFactor: Computation of Bayes factors for common designs* [R package version 0.9.12-4.3]. https://CRAN.R-project.org/package=BayesFactor

Mulder, J. (2016). Bayes factors for testing order–constrained hypotheses on correlations. *Journal of Mathematical Psychology*, *72*, 104–115.

Murphy, K. P. (2007). *Conjugate Bayesian analysis of the Gaussian distribution* (tech. rep.). University of British Columbia. New York, Springer.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

Nathoo, F. S., & Masson, M. E. J. (2016). Bayesian alternatives to null–hypothesis significance testing for repeated–measures designs. *Journal of Mathematical Psychology*, *72*, 144–157.

Nelder, J. (1977). A reformulation of linear models. *Journal of the Royal Statistical Society: Series A (General)*, *140*(1), 48–63.

Ng, K. W., Tian, G.-L., & Tang, M.-L. (2011). Dirichlet and related distributions: Theory, methods and applications.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Wiley.

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226.

Oberauer, K. (2022). The importance of random slopes in mixed models for bayesian hypothesis testing. *Psychological Science*, *33*(4), 648–665.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements: Eliciting Experts' Probabilities*. John Wiley & Sons.

O'Hara, R. B., Sillanpää, M. J., et al. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian analysis*, *4*(1), 85–117.

Olive, D. J. (2017). *Linear regression*. Springer International Publishing.

Paré, G., Cook, N. R., Ridker, P. M., & Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the Women's Genome Health Study. *PLoS Genetics*, *6*(6), e1000981.

Peña, V. (2018). *Bayesian model uncertainty and foundations*. [Doctoral dissertation, Duke University] [Retrieved from https://hdl.handle.net/10161/17494].

Petrocelli, J. V. (2003). Hierarchical multiple regression in counseling research: Common problems and possible remedies. *Measurement and Evaluation in Counseling and Development*, *36*(1), 9–22.

Pfister, R. (2021). Variability of bayes factor estimates in bayesian analysis of variance. *The Quantitative Methods for Psychology*, *17*(1), 40–45.

Phillips, P. C. B. (1988). The characteristic function of the Dirichlet and multivariate f distributions. *Cowles Foudation for Research in Econonmics*, 1–17.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., Carter, E. C., et al. (2019). Predicting human decisions with behavioral theories and machine learning. *arXiv preprint arXiv:1904.06866*.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. *Thirty-First AAAI Conference on Artificial Intelligence*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing*.

Porwal, A., & Raftery, A. E. (2022). Comparing methods for statistical inference with model uncertainty. *Proceedings of the National Academy of Sciences*, *119*(16), e2120737119.

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single-and dual-process models of recognition memory. *Journal of Mathematical Psychology*, *55*, 36–46.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in the assessment of ROC asymmetries. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 224–232.

Quintana, F. A. (2006). A predictive view of Bayesian clustering. *Journal of Statistical Planning and Inference*, *136*(8), 2407–2429.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, *92*(437), 179–191.

Rao, C. (2009). Multiple comparison procedures-a note and a bibliography. *Journal of Statistics*, *16*(1), 66–109.

Rasmussen, C. E., et al. (1999). The infinite Gaussian mixture model. *NIPS*, *12*, 554–560.

Reich, B. J., & Ghosh, S. K. (2019). *Bayesian statistical methods*. Chapman & Hall/CRC.

Robert, C. P. (2016). The expected demise of the Bayes factor. *Journal of Mathematical Psychology*, *72*, 33–37.

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—crisis and resolution in statistical inference. *The American Statistician*, *73*, 243–252.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist*, *88*(2), 313–338.

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2016). Model comparison in ANOVA. *Psychonomic Bulletin & Review*, *23*, 1779–1786.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, *25*, 102–113.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*, 877–903.

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

Rouder, J. N., Morey, R. D., Verhagen, A. J., Swagman, A. R., & Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, *22*, 304–321.

Rouder, J. N., & Haaf, J. M. (2021). Are there reliable qualitative individual difference in cognition? *Journal of Cognition*, *4*(1).

Rouder, J. N., Schnuerch, M., Haaf, J. M., & Morey, R. D. (2022). Principles of model specification in ANOVA designs. *Computational Brain & Behavior*.

Rousseau, J. (2007). Approximating interval hypothesis: P-values and Bayes factors. In J. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. Smith, & M. West (Eds.), *Bayesian statistics 8: Proceedings of the eighth Valencia international meeting june 2–6, 2006* (pp. 417–452, Vol. 8). Oxford University Press.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence-inconsistent implicit and explicit attitudes. *Psychological Science*, *17*(11), 954–958.

Saks, M. J., Hollinger, L. A., Wissler, R. L., Evans, D. L., & Hart, A. J. (1997). Reducing variability in civil jury awards. *Law and Human Behavior*, *21*(3), 243–256.

Savalei, V., & Dunn, E. (2015). Is the call to abandon $p-$values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142.

Schuringa, E., Spreen, M., & Bogaerts, S. (2014). Inter-rater and test-retest reliability, internal consistency, and factorial structure of the instrument for forensic treatment evaluation. *Journal of Forensic Psychology Practice*, *14*(2), 127–144.

Schuringa, E., Spreen, M., & Bogaerts, S. (2019). Inpatient violence in forensic psychiatry: Does change in dynamic risk indicators of the IFTE help predict short term inpatient violence? *International journal of law and psychiatry*, *66*, 101448.

Schuringa, E., Spreen, M., & Bogaerts, S. (2021). Treatment evaluation in forensic psychiatry. which one should be used: The clinical judgment or the instrument-based assessment of change? *International journal of offender therapy and comparative criminology*, 0306624X211023921.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, *136*, 2144–2162.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical–Bayes multiplicity adjustment in the variable–selection problem. *The Annals of Statistics*, *38*, 2587–2619.

Segal, D. L. (2010). Diagnostic and statistical manual of mental disorders (dsm-iv-tr). *The Corsini Encyclopedia of Psychology*, 1–3.

Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*, 1953–1967.

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of ρ values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71.

Sheather, S. (2009). *A modern approach to regression with R*. New York, Springer.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Sholts, S. B., Flores, L., Walker, P. L., & Wärmländer, S. K. (2011). Comparison of coordinate measurement precision of different landmark types on human crania using a 3D laser scanner and a 3D digitiser: implications for applications of digital morphometrics. *International Journal of Osteoarchaeology*, *21*(5), 535–543.

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., Gronau, Q. F., Kalish, M. L., McMullin, S. D., Navarro, D. J., & Shiffrin, R. M. (2022). Statistics in the service of science: Don't let the tail wag the dog. *Computational Brain & Behavior*.

Spreen, M., Brand, E., Ter Horst, P., & Bogaerts, S. (2014, November). *Handleiding en methodologische verantwoording hkt-r, historisch, klinische en toekomstige–revisie*. Dr. van Mesdag kliniek.

Stan Development Team. (2019). RStan: The R interface to Stan [R package version 2.19.2]. http://mc-stan.org/

Stan Development Team. (2022). *Stan modeling language users guide and reference manual, 2_30*. https://mc-stan.org

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of roc functions: A diffusion model analysis. *Journal of memory and language*, *70*, 36–52.

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Swets, J. A. (1986). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, *99*, 181–198.

Tanner Jr., W. P., & Swets, J. A. (1954). A decision–making theory of visual detection. *Psychological Review*, *61*, 401–409.

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075.

Teh, Y. W. (2010). Dirichlet Process.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Tijmstra, J. (2018). Why checking model assumptions using null hypothesis significance tests does not suffice: A plea for plausibility. *Psychonomic Bulletin & Review*, *25*, 548–559.

Trotta, R. (2008). Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, *49*(2), 71–104.

Tuente, S. K., Bogaerts, S., & Veling, W. (2021). Mapping aggressive behavior of forensic psychiatric inpatients with self-report and structured staff-monitoring. *Psychiatry research*, *301*, 113983.

US Food and Drug Administration. (1988). Guideline for the format and content of the clinical and statistical sections of an application. *Rockville, MD: US Food and Drug Administration.*

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without “p-ing” everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, *45*(3), 1–67.

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients’ mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

van den Bergh, D., Clyde, M. A., Komarlu Narendra Gupta, A. R., de Jong, T., Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2021a). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, 1–21.

van den Bergh*, D., & Dablander*, F. (2022). Flexible Bayesian multiple comparison adjustment using Dirichlet process and beta-binomial model priors. *Manuscript submitted for publication. *shared authorship.*

van den Bergh, D., Schuringa, E., & Wagenmakers, E.-J. (2023). Augmenting predictive models in forensic psychiatry with cultural consensus theory. *Manuscript submitted for publication.* 10.31234/osf.io/9kp3y

van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology.*, *120*(1), 73–96.

van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (in press). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *Behavior Research Methods.*

van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021b). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245921992035.

van der Schoot, R., Winter, S., Ryan, O., Zondervan–Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 300–319). Oxford University Press.

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (Eds.). (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, *25*, 1–4.

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2021). Bayes factors for mixed models. *Computational Brain & Behavior.*

van Doorn, J., Aust, F., Haaf, J. M., Stefan, A. M., & Wagenmakers, E.-J. (2022). Bayes factors for mixed models: Perspective on responses. *PsyArXiv.*

van Doorn, J., Haaf, J. M., Stefan, A. M., Wagenmakers, E.-J., Cox, G. E., Davis-Stober, C., Heathcote, A., Heck, D. W., Kalish, M., Kellen, D., et al. (2022). Bayes factors for mixed models: A discussion. *PsyArXiv.*

van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 1–14.

Venables, W. N. (2000). Exegeses on linear models. http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf

Vess, J. (2001). Development and implementation of a functional skills measure for forensic psychiatric inpatients. *The Journal of Forensic Psychiatry*, *12*(3), 592–609.

Vovk, V. G. (1993). A logic of probability, with application to the foundations of statistics. *Journal of the Royal statistical society: series B (Methodological)*, *55*(2), 317–341.

Wagenmakers, E.-J., Kucharský, Š., & the JASP Team. (2020). *The JASP data library*. JASP Publishing. https://psyarxiv.com/vr2u8/

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*, 192– 196.

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Westfall, P. H., Johnson, W. O., & Utts, J. M. (1997). A Bayesian perspective on the Bonferroni adjustment. *Biometrika*, *84*, 419–427.

Westfall, P. H. (1997). Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, *92*(437), 299–306.

Wickens, T. (2001). *Elementary signal detection theory*. Oxford University Press.

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). Bayesian model search and multilevel inference for snp association studies. *The annals of applied statistics*, *4*(3), 1342.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1), 1–17.

Wrinch, D., & Jeffreys, H. (1919). On some aspects of the theory of probability. *Philosophical Magazine*, *38*, 715–731.

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, *42*, 369–390.

Yu, C., & Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, *105*(2), 319–335.

Yu, C.-H., Prado, R., Ombao, H., & Rowe, D. (2018). A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fmri. *Journal of the American Statistical Association*, *113*(524), 1395–1410.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g–prior distributions. In P. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233–243). North–Holland.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística y de Investigación Operativa*, *31*(1), 585–603.

13

Nederlandse Samenvatting

Kenmerkend voor de empirische wetenschap is het omgaan met onzekerheid. Onzekerheid kent vele bronnen en vindt zijn weg naar empirisch onderzoek op zowel verwachte als onverwachte manieren. Stel bijvoorbeeld dat een wetenschapper het mentale welzijn van een proefpersoon wil meten met een vragenlijst. Ze stelt de proefpersoon een reeks vragen en de proefpersoon geeft eerlijk antwoord. In deze situatie zijn de antwoorden van de proefpersoon, dat wil zeggen de data, ruizige metingen van datgene waarin de wetenschapper geïnteresseerd is, namelijk het welzijn van de proefpersoon. Om met deze onzekerheid rekening te houden is het gebruikelijk dat wetenschappers meerdere vragen stellen en aannemen dat de ruis in de antwoorden op individuele vragen verdwijnt bij het middelen. Dit is een bekende bron van onzekerheid, ook wel bekend als meetfout. Een minder bekende bron van onzekerheid is afkomstig van statistische modellen en de beslissingen die tussentijds worden genomen bij het uitvoeren van statistische analyses. In statistische modellen en de bijbehorende analyses worden bijvoorbeeld vaak aannames gedaan. Deze aannames kunnen echter worden geschonden en daarom hangen de conclusies van die analyses af van de (on)zekerheid waarmee aan de aannames wordt voldaan. In de empirische praktijk zien we vaak dat dit resulteert in een tweestaps-procedure. Eerst wordt onderzocht of de aannames van een volgende test worden geschonden. Worden deze aannames niet geschonden, dan wordt de test van belang uitgevoerd. Deze tweestaps-procedure begaat echter een subtiele fout. De door de tweede test gerapporteerde onzekerheid houdt geen rekening met de (on)zekerheid waarmee de eerste test heeft bepaald of de aannames zijn geschonden. In feite wordt de onzekerheid van de eerste test volledig genegeerd, waardoor de tweede test overmoedige resultaten oplevert. In het algemeen bestaat de data analyse in empirische studies vaak uit meerdere tussenstappen en tussentijdse beslissingen. Elk van deze stappen gaat gewoonlijk gepaard met onzekerheid en vrijwel geen enkele tussentijdse beslissing wordt met absolute zekerheid genomen. In veel studies wordt de onzekerheid van de verschillende tussenstappen echter genegeerd, waardoor een vals betrouwbaarheidsniveau ontstaat dat de conclusies kan beïnvloeden. Het centrale thema van dit proefschrift is de onzekerheid die ontstaat bij het uitvoeren van multi-stap inferentie en hoe alle onzekerheid in de uiteindelijke resultaten kan worden verwerkt.

13.1 Onzekerheid Binnen een Model

Het eerste deel van dit proefschrift gaat over onzekerheid binnen een statistisch model. Dit deel illustreert dat beslissingen die voorafgaan aan de toepassing van een statistisch model of analyse onzekerheid kunnen verbergen. Stel dat een groep patiënten met een psychische stoornis door vijf psychiaters wordt gescoord op verschillende items, zoals psychotisch gedrag, impulsief gedrag en probleeminzicht. Een gebruikelijke eerste stap bij het analyseren van dergelijke data is het nemen van het steekproefgemiddelde van alle vijf de psychiaters om één score voor elke patiënt op elk item te verkrijgen. In een tweede stap worden deze gemiddelden onderworpen een statistische test. Deze tweestapsprocedure gaat voorbij aan een belangrijke bron van onzekerheid, namelijk dat de verschillende psychiaters niet allemaal dezelfde score gaven. Om precies te zijn, door het naïef middelen van de scores wordt impliciete een aanname gemaakt dat de vijf psychiaters uitwisselbaar zijn en dat hun individuele verschillen irrelevant zijn. Doorgaans is deze veronderstelling ongegrond omdat verschillende psychiaters verschillende achtergronden hebben en zinvolle individuele verschillen kunnen hebben die resulteren in heterogeen scoringsgedrag. Maar door deze bron van onzekerheid onder het tapijt te vegen, wordt de variabiliteit in de antwoorden genegeerd. Dit leidt tot een te groot vertrouwen in wat nu de "waargenomen" data zijn (d.w.z. de gemiddelde scores) die in latere analyses worden gebruikt. Om te zien waarom dit leidt tot een opgeblazen vertrouwen, merk op dat deze procedure op zijn minst de standaardfouten negeert die samengaan met de steekproefgemiddelden. Daarom zijn de waarnemingen die door latere analyses worden gebruikt variabeler dan de analyses weten. Maar omdat alle latere analyses blind zijn voor deze variabiliteit, zijn de onzekerheidsintervallen kleiner dan ze zouden moeten zijn, wat leidt tot overmoedige conclusies.

Het doel van het eerste deel van dit proefschrift was om modellen te ontwikkelen die expliciet rekening houden met het feit dat verschillende individuen de scores hebben gegeven en om de onzekerheid goed te kwantificeren. Hoewel de in dit deel geanalyseerde data zich richten op patiënten in forensisch psychiatrische ziekenhuizen en dus vrij gespecialiseerd zijn, is de structuur van de geanalyseerde data vrij gebruikelijk in de empirische wetenschap. Bijvoorbeeld, data verzameld via vragenlijsten die herhaaldelijk door deelnemers worden ingevuld of data in de onderwijspsychologie waarbij opstellen of andere producten door meerdere beoordelaars worden gescoord, hebben vaak een soortgelijke structuur. Als zodanig zijn de in dit deel ontwikkelde methoden generaliseerbaar naar andere toepassingen buiten een forensische setting.

13.2 Onzekerheid Tussen Meerdere Modellen

Het tweede deel van dit proefschrift gaat over onzekerheid wanneer er meerdere modellen in het spel zijn. Stel dat we het risico op een geweldsuitbarsting in het cohort van onze patiënten met een psychische stoornis willen voorspellen. Met behulp van alle beschikbare data willen we een model construeren dat het risico op een geweldsuitbarsting in de toekomst nauwkeurig voorspelt. Als we echter naïef alle door psychiaters gescoorde items en andere achtergrondvariabelen opnemen, lopen we het risico op overfitting en kan ons model toekomstige geweldsuitbarstingen slecht voorspellen. Een gebruikelijke aanpak om overfitting tegen te gaan is het gebruik van een tweestaps-procedure, waarbij in de eerste stap één model wordt geselecteerd en in de tweede stap dat model wordt geïnterpreteerd en gebruikt voor voorspellingen. Deze procedure in twee stappen gaat echter opnieuw systematisch voorbij aan de onzekerheid en kan leiden tot overmoedige conclusies. Er bestaat grote onzekerheid over welk model het “beste” model is dat in de tweede stap moet worden gebruikt. Soms is er niet een enkel best model dat superieur is aan alle andere kandidaatmodellen. In plaats daarvan is er vaak een overvloed aan modellen die adequate voorspellingen doen en redelijke verklaringen bieden voor de gegevens. Maar door één model te kiezen, negeren we deze modelonzekerheid en doen we alsof we het enige echte model hebben ontdekt waarop we onze conclusies kunnen baseren. Daardoor worden we overmoedig en overschatten we de grootte van de vastgestelde effecten (bijv. Hoeting et al., 1999; Porwal and Raftery, 2022, Chapter 5)

Het tweede deel van dit proefschrift had tot doel nieuwe statistische methoden te ontwikkelen om de onzekerheid over verschillende modellen te kwantificeren. De belangrijkste benadering die centraal staat in dit deel is *model averaging*. In plaats van een enkel model te selecteren voor voorspellingen, maken we voorspellingen met behulp van alle beschouwde modellen en wegen we de voorspellingen van elk model naar hun relatieve plausibiliteit in het licht van de data.

13.3 Onzekerheid Aanvaarden voor Iedereen

Het derde deel van dit proefschrift gaat over het toegankelijk maken van modelmiddeling voor iedereen zonder wiskundige achtergrond en programmeerkennis. Een groot deel van de statistische literatuur houdt zich bezig met de ontwikkeling van nieuwe en belangrijke methoden. Maar de weg om deze methoden in de praktijk te brengen is echter vol obstakels, als er al een begaanbaar pad te zien is. Geïnteresseerden kunnen bijvoorbeeld de wiskundige achtergrond missen om de afleidingen te begrijpen of de programmeerkennis om een nieuwe techniek toe te passen. Deze belemmeringen maken het moeilijk om nieuwe ontwikkelingen in praktijk te brengen. Het derde deel reflecteert op de literatuur over Bayesiaanse modelvergaring en richtte zich op het aanpassen, verfijnen en ontwikkelen van modelvergaringstechnieken aan statistische paradigma's die relevant zijn voor de psychologische praktijk. Dit werd gedaan door de technieken te implementeren in het gratis en open-source statistische softwareprogramma JASP (JASP Team, 2022). Hierdoor is het mogelijk om de in de eerste twee delen van het proefschrift ontwikkelde ideeën en technieken in de praktijk te brengen zonder dat daarvoor een diepgaande wiskundige achtergrond of geavanceerde programmeerkennis nodig is.

13.4 Slotopmerkingen

De centrale les van dit proefschrift is dat analyses die uit meerdere stappen bestaan moeten worden vermeden. Kenmerkend voor multi-step inferentie is dat onzekerheid in eerdere stappen gemakshalve wordt vergeten in volgende stappen. Het verwaarlozen van deze onzekerheid leidt uiteindelijk tot overmoedige inferentie. In plaats van de onzekerheid van eerdere stappen te vergeten, moet deze onzekerheid worden omarmd en moet in de uiteindelijke conclusies rekening worden gehouden de onzekerheid van alle stappen. Dit proefschrift introduceerde nieuwe methoden en verbeterde de toegankelijkheid van oudere methoden. Ik hoop dat met dit proefschrift de praktijk van het negeren van onzekerheid door verschillende inferentiestappen samen te voegen een overblijfsel uit het verleden wordt en dat toekomstige studies de onzekerheid van de afzonderlijke stappen omarmen door multi-model inferentie toe te passen.

14 Acknowledgements

The last few years have been an incredible adventure. I consider myself incredibly lucky to have met so many people without whom this experience would not have been nearly as interesting and fun.

First of all, I would like to thank my supervisors, without whom my PhD would definitely have derailed early on. EJ, thank you for all your patience with my sloppy introductions. Our meetings were always a success, whether we were actually engaged in the topic of our meeting, or got sidetracked with JASP and chess.

Maarten, thank you for your help deriving Gibbs samplers and your practical guidance throughout. Whenever I encountered what I thought was an insurmountable obstacle (e.g., it will take until the heat death of the universe before these MCMC chains converge), you always found a solution to get me back on track and continue to do so until today.

Alexander, thank you for all the mathematical help and your repeated emphasis to also focus on furthering my own development. Without you, Fabian and I would still be stuck in the first book on real analysis and the paper on comparing variances would still be a draft.

I'd like to thank my paranymphs, who have supported me throughout my journey and have always been there for me when I needed them.

Fabian, your perseverance and motivation are impressive and admirable. During our joint projects, your enthusiasm always motivated me to dig deeper into whatever we were investigating. For example, without your many sidequests, I would never have read anything about causal inference and differential equations, nor would I have done a project in Julia.

Alexandra, aside from your academic brilliance, you possess unparalleled social skills. This was already clear to me during our masters, where you effortlessly brought together different social groups from Tübingen and Amsterdam on several occasions (e.g., in Liège, Vienna, and Prague). Without you, I would not have met Selina, I would not have gotten to know so many interesting people, and the whole PhD journey would not have been nearly as much fun.

Without a doubt, the most enjoyable after-hours adventures were with ”The Gang.” I will always remember the cozy lunches, movie evenings, game nights, Crea drinks, and the many pizzas (RIP Torino's) and other dinners.

The most fun side-quest during my PhD was undoubtedly my work on JASP. It was usually very easy to put the manuscripts aside for a while and instead fix some bugs in JASP or add some new functionality. This was of course accompanied by nice discussions and conversations with among others, Tim de Jong, Joris, Frans, Bruno, Rens, Koen, Šimon, František, Erik-Jan van Kesteren, Jonas, Johnny, and Tim Draws. Without the JASP team, I would probably never have learned so much about R, C++, and cross-platform programming, and their many terrible hidden subtleties.

Finally and probably most importantly, I'd like to thank my family. Pap, mam, zonder jullie onuitputbare hulp was ik nooit gekomen waar ik nu ben. Dank jullie wel voor jullie oneindige steun. Pap, ik vindt het erg leuk dat we toch nog samen een artikel hebben geschreven. Mam, dankjewel voor al je hulp en steun door de jaren heen. Bij het minste of geringste bood je altijd aan om te helpen (bv als ik 's nachts een lekke band had), en zelfs nu sta je nog altijd voor me klaar.

Mattis, het was altijd gezellig om met jou iets in R te programmeren of aan een wiskunde opdracht te werken en daar heb ik ook veel van geleerd. Al tijdens mijn bachelor introduceerde je mij LaTeX en ook was de eerste keer dat ik iets over de programmeertaal Julia hoorde via jou. Naast de studie was het ontzettend leuk om samen stoom af te blazen, in het roze legioen te voetballen, en is het altijd gezellig om met jou, Dorina en Selina, toe te kijken hoe Lasse, Aiko, en Frea de wereld ontdekken.

Helen & Peter, al hebben jullie niet direct iets bijgedragen aan de inhoud van dit boek, wil ik jullie toch bedanken voor alle wijze levenslessen die ik van jullie heb geleerd. Als het gaat om klussen, feestjes plannen, of naar het buitenland emigreren (mocht dat ooit gebeuren), dan zal ik jullie altijd om advies vragen.

Frea, dankjewel voor het opvrolijken van de laatste fase van mijn PhD. Jouw schattigheid, capaciteit om mij van mijn werk te houden, en de capriolen waarmee je een lach op mijn gezicht brengt zijn ongeëvenaard in deze wereld. Ten minste, totdat je in april een grote zus wordt.

Selina, thank you for your unwavering support, love, and encouragement throughout these past years. After a day of work you would always be up for something fun, painting together, going to the movies, raising pet rats, renovating something in our place from our never-ending list, or just relaxing on the couch together. Whenever I felt down for some reason, you would find a way to cheer me up and stop me from sulking about it. Your ability to focus on the important things in life is something I truly admire about you. Thank you for keeping me in check and thank you for all the wonderful time we spent together.

15
Contributions

15.1 Part I: Cultural Consensus Theory

chapter 2: Parsimonious Estimation of Signal Detection Models from Confidence Ratings

This chapter is published as: Selker, R., van den Bergh, D., Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*, 1953–1967.

RS, AHC, and EJW proposed the study. RS wrote the first draft of the manuscript and implement the initial models. DvdB analyzed the data and finalized the manuscript. EJW and AHC provided detailed feedback on the manuscript and guidance throughout.

chapter 3: Cultural Consensus Theory for the Evaluation of Patients' Mental Health Scores in Forensic Psychiatric Hospitals

This chapter is published as: van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

DvdB, SB, MS, RF, JV, WHB, and EJW proposed the study. DvdB implemented the software, conducted the simulation study, and wrote the first draft of the manuscript. SB, MS, RF, JV, and EJW provided detailed feedback on the manuscript and guidance throughout.

Chapter 4: Augmenting Predictive Models in Forensic Psychiatry with Cultural Consensus Theory

This chapter is submitted as: van den Bergh, D., Schuringa, E., & Wagenmakers, E.-J. (2023). Augmenting predictive models in forensic psychiatry with cultural consensus theory. *Manuscript submitted for publication.* 10.31234/osf.io/9kp3y

DvdB, and EJW proposed the study. ES collected, shared, and anonymized the data. DvdB implemented the software, analyzed the data, and wrote the first draft of the manuscript. ES and EJW provided detailed feedback on the manuscript and guidance throughout.

15.2 Part II: Bayesian Model Averaging

Chapter 5: A Cautionary Note on Estimating Effect Size

This chapter is published as: van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021b). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245921992035.

DvdB, JMH, AL, JNR, and EJW proposed the study. DvdB and JMH analzyed the data. JNR provided the code used in the analyses. DvdB and JMH wrote the first draft of the manuscript. JMH, AL, JNR, and EJW provided detailed feedback on the manuscript and guidance throughout.

Chapter 6: Default Bayes Factors for Testing the (In)equality of Several Population Variances

This chapter is published as: Dablander*, F., van den Bergh*, D., Wagenmakers, E.-J., & Ly, A. (in press). Default Bayes factors for testing the (in)equality of several population variances. *Bayesian Analysis. *shared authorship*

FD and DvdB proposed the study. They both worked out the initial derivations and proofs for the deterministic K = 2 case with the help of AL. FD wrote the first draft of the manuscript and analyzed the data. FD developed the software package with the help of DvdB. AL extended the results to the K ≥ 2 case and provided the proofs shown in Appendices E.2 and E.2.2. FD, DvdB, and AL wrote the manuscript. EJW provided detailed feedback on the manuscript and guidance throughout.

chapter 7: Flexible Bayesian Multiple Comparison Adjustment Using Dirichlet Process and Beta-Binomial Model Priors

This chapter is submitted as: van den Bergh*, D., & Dablander*, F. (2022). Flexible Bayesian multiple comparison adjustment using Dirichlet process andbeta-binomial model priors. *Manuscript submitted for publication.* https://arxiv.org/abs/2208.07086. **shared authorship*

DvdB and FD proposed the study. They both worked out the initial derivations and designed the simulation study. FD wrote the first draft of the manuscript. DvdB derived the prediction rule for the Beta-binomial, developed the software package and analyzed the data. DvdB and FD wrote the manuscript.

15.3 Part III: JASP

chapter 8: A Tutorial on Bayesian Multi-Model Linear Regression with BAS and JASP

This chapter is published as: van den Bergh, D., Clyde, M. A., Komarlu Narendra Gupta, A. R., de Jong, T., Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2021a). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, 1–21.

DvdB and EJW proposed the study. MAC implemented the R package BAS. DvdB, ARKNG, TdJ, and QFG implemented the functionality of the R package BAS in JASP. DvdB analyzed the data and wrote the initial draft of the manuscript. All authors provided detailed feedback on the manuscript. AL and EJW provided guidance throughout.

Chapter 9: A Tutorial on Conducting and Interpreting a Bayesian ANOVA in JASP

This chapter is published as: van den Bergh, D., van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology.*, *120*(1), 73–96.

DvdB and EJW proposed the study. DvdB, JvD, MM, TD, EJvK, KD, QFG, SK, ARKNG, ASa, JGV, and AL implemented the functionality in JASP. DvdB analyzed the data examples and wrote the initial draft of the manuscript. All authors provided detailed feedback on the manuscript. MM, AL, and EJW provided guidance throughout.

Chapter 10: Bayesian Repeated-measures ANOVA: An updated Methodology Implemented in JASP

This chapter is published as: van den Bergh, D., Wagenmakers, E.-J., & Aust, F. (in press). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *Behavior Research Methods.*

DvdB, EJW and FA proposed the study. FA worked out the discrepancy between the frequentist and Bayesian ANOVA. DvdB implemented the new model specifications in JASP and analyzed the data. DvdB and FA wrote the first draft of the manuscript. EJW and FA provided detailed feedback on the manuscript and guidance throughout.

16 Publications

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., **van den Bergh**, **D.**, & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366.

Barbalat, G., **van den Bergh**, **D.**, & Kossakowski, J. (2019). Outcome measurement in mental health services: Insights from symptom networks. *BMC Psychiatry*, *19*(1), 1–9.

Dablander*, F., **van den Bergh***, **D.**, Wagenmakers, E.-J., & Ly, A. (in press). Default Bayes factors for testing the (in) equality of several population variances. *Bayesian Analysis. *shared authorship.*

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., Mulder, J., Palfi, B., Schönbrodt, F. D., Tendeiro, J. N., **van den Bergh**, **D.**, van Lissa, C., van Ravenzwaaij, D., Vanpaemel, W., Wagenmakers, E.-J., William, D. R., Zondervan-Zwijnenburg, M., & Hoijtink, H. (2022). A review of applications of the Bayes factor in psychological research. *Psychological Methods.*

Hinne, M., Gronau, Q. F., **van den Bergh**, **D.**, & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215.

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Alexander, L., **van den Bergh**, **D.**, Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., Cheung, F., Cimpian, A., Dohle, S., Donnellan, M. B., Hahn, A., Hall, M. P., Jiménez-Leal, W., Johnson, D. J., Lucas, R. E., Monin, B., Montealegre, A., Mullen, E., Pang, J., Ray, J., Reinero, D. A., Reynolds, J., Sowden, W., Storage, D., Su, R., Tworek, C. M., Van Bavel, J. J., Walco, D., Wills, J., Xu, X., Yam, K. C., Yang, X., Cunningham, W. A., Schweinsberg, M., Urwitz, M., & Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*.

Ly, A., Stefan, A., van Doorn, J., Dablander, F., **van den Bergh, D.**, Sarafoglou, A., Kucharsky, S., Derks, K., Gronau, Q. F., Komarlu Narendra Gupta, A. R., Boehm, U., van Kesteren, E.-J., Hinne, M., Matzke, D., Marsman, M., & Wagenmakers, E.-J. (2020). The Bayesian methodology of sir Harold Jeffreys as a practical alternative to the p-value hypothesis test. *Computational Brain & Behavior.*, *3*(2), 153–161.

Ly, A., **van den Bergh, D.**, Bartoš, F., & Wagenmakers, E.-J. (2021). Bayesian inference with JASP. *The ISBA Bulletin*, *28*, 7–15. https://bayesian.org/wp-content/uploads/2021/03/2103.pdf

Maier, M., Bartoš, F., Quintana, D. S., Dablander, F., **van den Bergh, D.**, Marsman, M., Ly, A., & Wagenmakers, E.-J. (2022). Model-averaged Bayesian t-tests. *Manuscript submitted for publication.*

Pfadt, J. M., **van den Bergh, D.**, Sijtsma, K., & Wagenmakers, M., E.-J. Moshagen. (2022). A tutorial on Bayesian single-test reliability analysis with jasp. *Behavior Research Methods.*

Pfadt, J. M., **van den Bergh, D.**, & Moshagen, M. (2022). Classical and bayesian uncertainty intervals for the reliability of multidimensional scales. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–15.

Pfadt, J. M., **van den Bergh, D.**, Sijtsma, K., Moshagen, M., & Wagenmakers, E.-J. (2021). Bayesian estimation of single-test reliability coefficients. *Multivariate Behavioral Research*, *0*(0), 1–30.

Selker, R., **van den Bergh, D.**, Criss, A. H., & Wagenmakers, E.-J. (2019). Parsimonious estimation of signal detection models from confidence ratings. *Behavior Research Methods*, *51*, 1953–1967.

van den Bergh, D., Bogaerts, S., Spreen, M., Flohr, R., Vandekerckhove, J., Batchelder, W. H., & Wagenmakers, E.-J. (2020). Cultural consensus theory for the evaluation of patients' mental health scores in forensic psychiatric hospitals. *Journal of Mathematical Psychology*, *98*, 102383.

van den Bergh, D., Clyde, M. A., Komarlu Narendra Gupta, A. R., de Jong, T., Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2021). A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behavior Research Methods*, 1–21.

van den Bergh*, **D.**, & Dablander*, F. (2022). Flexible Bayesian multiple comparison adjustment using Dirichlet process and beta-binomial model priors. *Manuscript submitted for publication. *shared authorship.*

van den Bergh, **D.**, Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science.*

van den Bergh, **D.**, Schuringa, E., & Wagenmakers, E.-J. (2023). Augmenting predictive models in forensic psychiatry with cultural consensus theory. *Manuscript submitted for publication.* 10.31234/osf.io/9kp3y

van den Bergh, **D.**, Tuerlinckx, F., & Verdonck, S. (2019). DstarM: An R package for analyzing two-choice reaction time data with the D∗M method. *Behavior Research Methods.*

van den Bergh, **D.**, van Doorn, J., Marsman, M., Draws, T., van Kesteren, E.-J., Derks, K., Dablander, F., Gronau, Q. F., Kucharsky, S., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année Psychologique/Topics in Cognitive Psychology.*, *120*(1), 73–96.

van den Bergh, **D.**, Vandermeulen, N., Lesterhuis, M., De Maeyer, S., Van Steendam, E., Rijlaarsdam, G., & van den Bergh, H. (2022). How prior information from national assessments can be used when designing experimental studies without a control group. *Journal of Writing Research.*

van den Bergh, **D.**, Wagenmakers, E.-J., & Aust, F. (in press). Bayesian repeated-measures ANOVA: An updated methodology implemented in JASP. *Behavior Research Methods.*

van Doorn, J., **van den Bergh**, **D.**, Dablander, F., van Dongen, N., Derks, K., Evans, N. J., Gronau, Q. F., Haaf, J. M., Kunisato, Y., Ly, A., Marsman, M., Sarafoglou, A., Stefan, A., & Wagenmakers, E.-J. (2021). Strong public claims may not reflect researchers' private convictions. *Significance*, *18*, 44–45.

van Doorn, J., **van den Bergh**, **D.**, Böhm, U., Dablander, F., Derks, K., Draws, T., Evans, N. J., Gronau, Q. F., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Komarlu Narendra Gupta, A. R., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 1–14.

Verschuere, B., De Schryver, M., **van den Bergh**, **D.**, Wagenmakers, E.-J., & Meijer, E. (2021). Are dishonest politicians more likely to be reelected? a Bayesian view. *Proceedings of the National Academy of Sciences*, *118*(6).

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., **van den Bergh**, **D.**, Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*, 58–76.

Part V

Appendices

A
Appendix of Chapter 2

This appendix contains both the formal BUGS model definition and the graphical representation of the SDT threshold model and the hierarchical SDT threshold model. The R code that calls the BUGS code is available at https://osf.io/v3b76/. The model definition and graphical representation define all priors and relations between parameters and data. For more information on the BUGS modeling language and the graphical representation of these models, see Lee and Wagenmakers (2013).

A.1 SDT Threshold Model

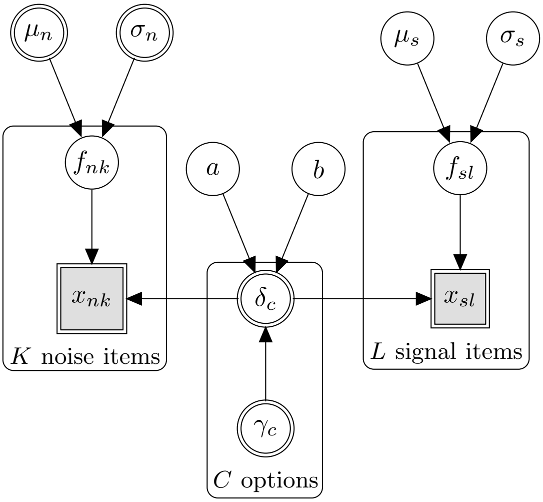


Figure A.1: Graphical model representation of the SDT threshold model.

Thresholds

$$\begin{aligned} a &\sim \text{Gamma}(0,5) \\ b &\sim \text{Gaussian}(0,3) \\ \gamma_c &\leftarrow \text{logit}\left(\frac{c}{C}\right) \\ \delta_c &\leftarrow a\gamma_c + b \end{aligned}$$

Signal

$$\begin{aligned} \mu_s &\sim \text{Uniform}(0,5) \\ \sigma_s &\sim \text{Uniform}(1,3) \\ f_{sl} &\sim \text{Gaussian}(\mu_s, 1/\sigma_s^2) \\ x_{sl} &\leftarrow \begin{cases} 1, & \text{if } f_{slk} \leq \delta_1 \\ c, & \text{if } \delta_{c-1} < f_{sl} \leq \delta_c \\ C, & \text{if } f_{sl} > \delta_{C-1} \end{cases} \end{aligned}$$

Noise

$$\begin{aligned} \mu_n &\leftarrow 0 \\ \sigma_n &\leftarrow 1 \\ f_{nk} &\leftarrow \text{Gaussian}(\mu_n, 1/\sigma_n^2) \\ x_{nk} &\leftarrow \begin{cases} 1, & \text{if } f_{nk} \leq \delta_1 \\ c, & \text{if } \delta_{c-1} < f_{nk} \leq \delta_c \\ C, & \text{if } f_{nk} > \delta_{C-1} \end{cases} \end{aligned}$$

A.2 Hierarchical SDT Threshold Model

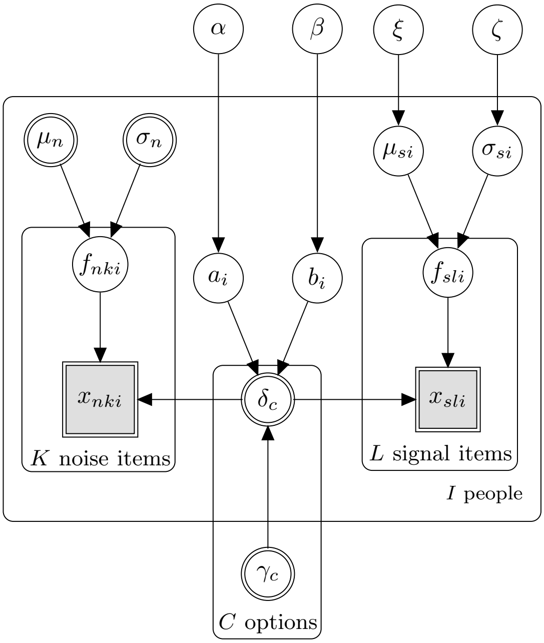


Figure A.2: Graphical model representation of the hierarchical SDT threshold model.

Hierarchical

$$\alpha \sim \text{Gaussian}(1,1)_{I(0,\text{inf})}$$
$$\beta \sim \text{Gaussian}(0,1)$$
$$\xi \sim \text{Gaussian}(1,0)_{I(0,\text{inf})}$$
$$\zeta \sim \text{Gaussian}(1.1,0)_{I(1,5)}$$

Thresholds

$$a_i \sim \text{Gaussian}(\alpha,1)$$
$$b_i \sim \text{Gaussian}(\beta,1)$$
$$\gamma_c \leftarrow \text{logit}\left(\frac{c}{C}\right)$$
$$\delta_{ci} \leftarrow a_i\gamma_c + b_i$$

Signal

$$\mu_{si} \sim \text{Gaussian}(\xi,1)$$
$$\sigma_{si} \sim \text{Gaussian}(\zeta,1)$$
$$f_{sli} \sim \text{Gaussian}(\mu_{si},1/\sigma_{si}^2)$$
$$x_{sli} \leftarrow \begin{cases} 1, & \text{if } f_{slk} \leq \delta_1 \\ c, & \text{if } \delta_{c-1} < f_{sl} \leq \delta_c \\ C, & \text{if } f_{sl} > \delta_{C-1} \end{cases}$$

Noise

$$\mu_n \leftarrow 0$$
$$\sigma_n \leftarrow 1$$
$$f_{nki} \leftarrow \text{Gaussian}(\mu_n,1/\sigma_n^2)$$
$$x_{nki} \leftarrow \begin{cases} 1, & \text{if } f_{nk} \leq \delta_1 \\ c, & \text{if } \delta_{c-1} < f_{nk} \leq \delta_c \\ C, & \text{if } f_{nk} > \delta_{C-1} \end{cases}$$

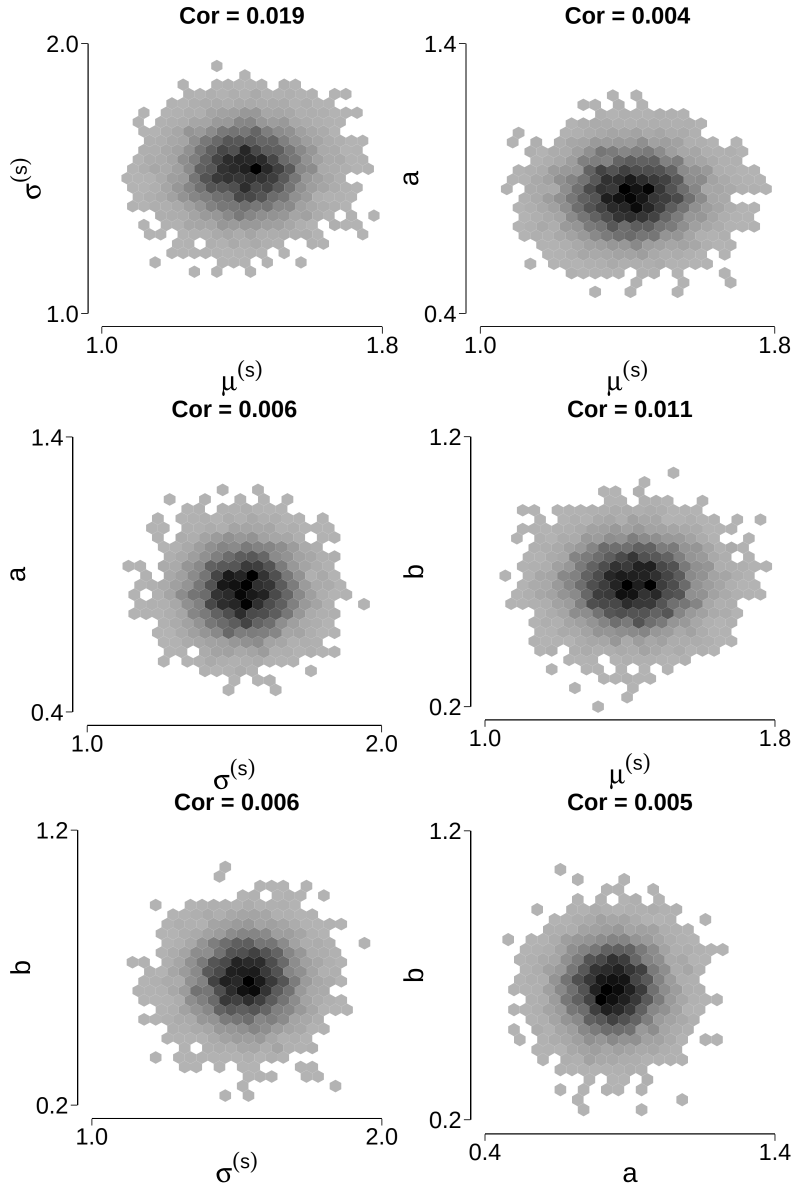


Figure A.3: Bivariate hex plots of the group-level parameters. A brighter color indicates a higher frequency of samples. The Pearson correlation between the posterior samples is shown on top of each panel. Note the negligible trade-off between the parameters.

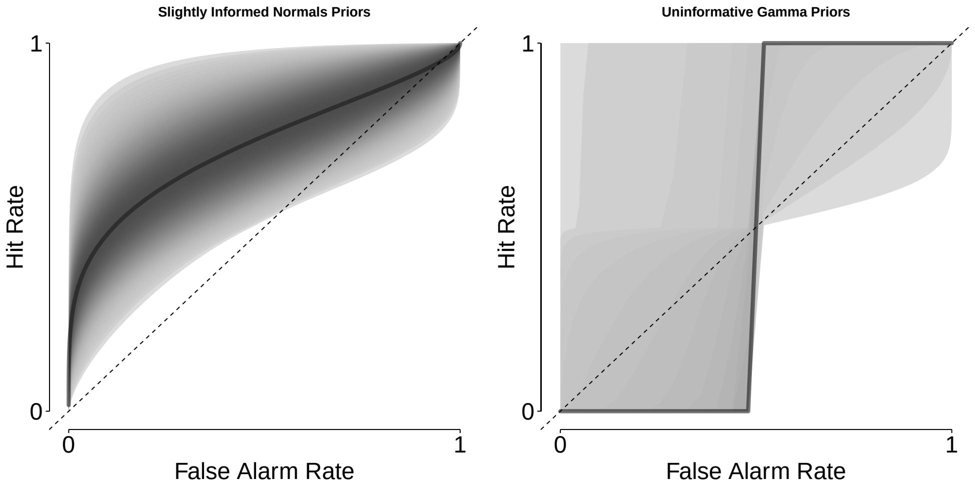


Figure A.4: Prior predictive ROCs for the proposed priors (left panel; see Figure A.2 for the priors) versus the standard uninformative gamma priors (right panel; $\alpha, \beta, \xi, \zeta \sim \text{Gamma}(0.01, 0.01)$).

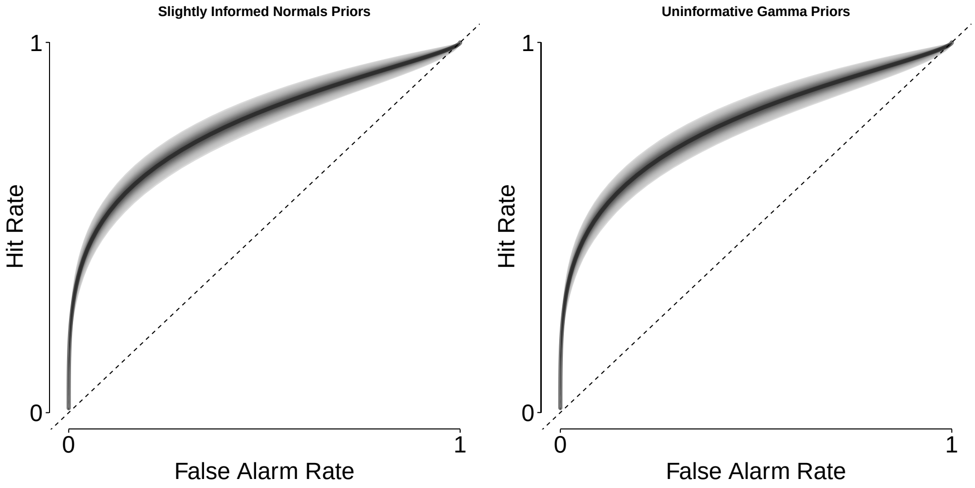


Figure A.5: Posterior predictive ROCs for the proposed priors (left panel; see Figure A.2 for the priors) versus the standard uninformative gamma priors (right panel; $\alpha, \beta, \xi, \zeta \sim \text{Gamma}(0.001, 0.001)$).

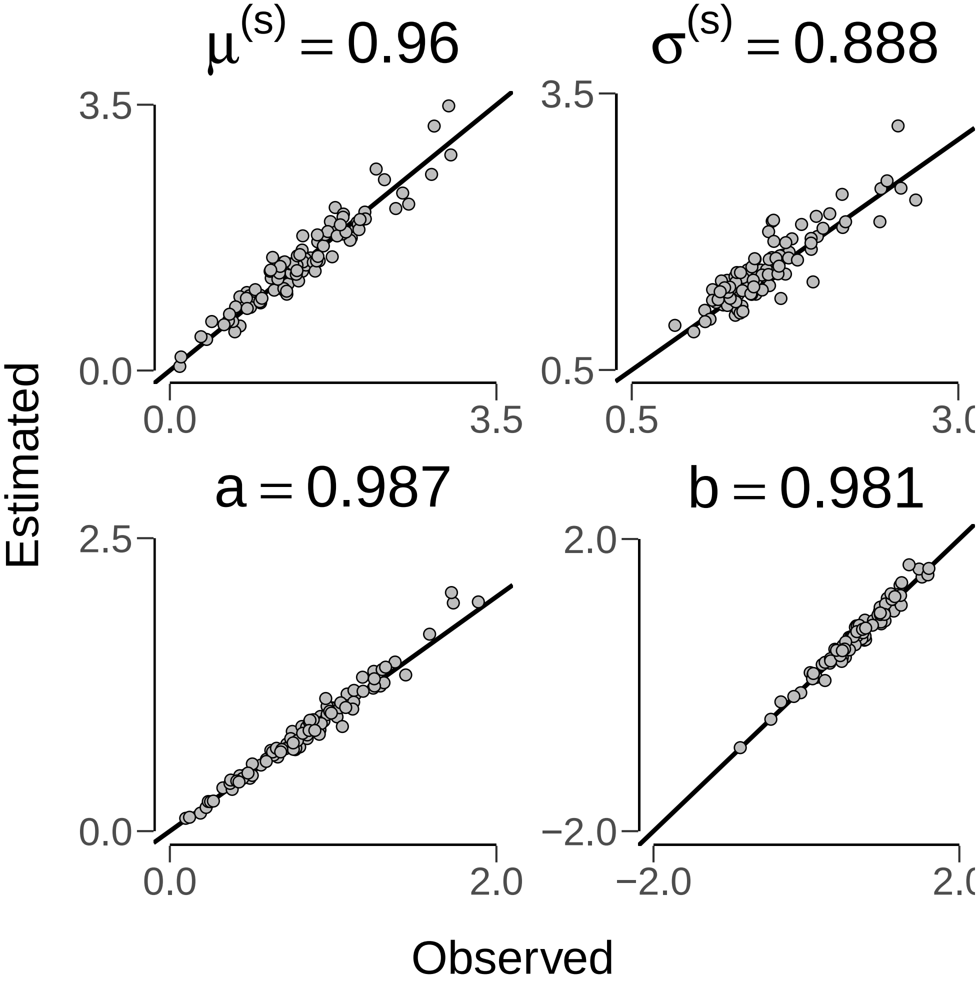


Figure A.6: Parameter retrieval of the group level parameters of the simulation study with the hierarchical model.

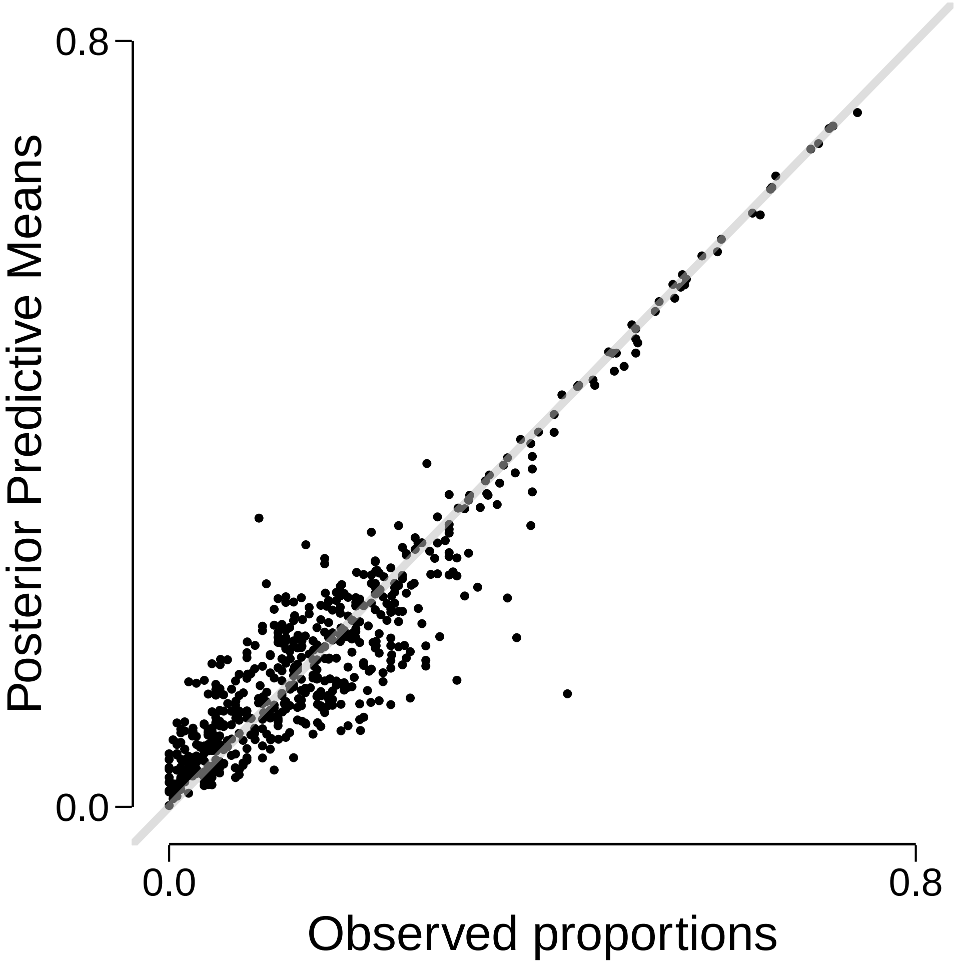


Figure A.7: Posterior predictive check for the data from Pratte et al. (2010). Observed proportions of a rating per person (x-axis) versus posterior predictive means of the model (y-axis). The model fits ratings with a higher observed proportion better than those with a lower observed proportion. This occurs because those ratings constitute more observations and are weighed more by the likelihood. Lower proportions are captured less well by the model. Likewise, the lower proportions are based on less data and are therefore more noisy.

B
Appendix of Chapter 3

B.1 Example Analysis

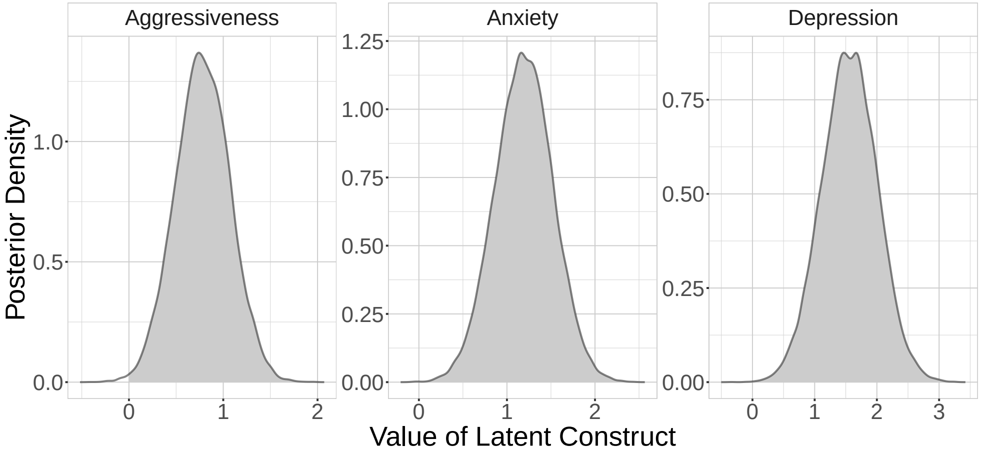


Figure B.1: Approximate posterior densities for the differences in latent constructs of two fictitious patients with response pattern. The probability that the difference is larger than 0 is above 0.99 for all constructs.

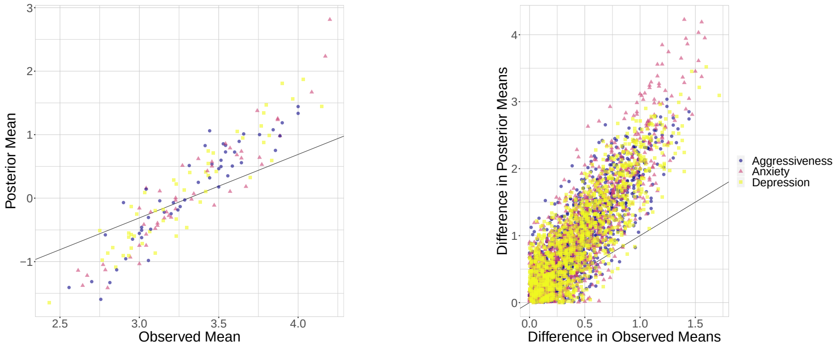


Figure B.2: The left panel plots the means of the observed ratings against the posterior means of the latent variables. The right panel shows for each combination of patients i, j the absolute difference in means, $|\hat{x}_i - \hat{x}_j|$, against the absolute difference in posterior means of the latent variables, $|\hat{\eta}_i - \hat{\eta}_j|$. Note that in the left panel, there is a difference in intercept because the responses are on a scale from 1 to 5, whereas the latent variables are assumed to have a mean of 0. The large spread in the right panel demonstrates that the sample mean is an unreliable indicator of the truth underlying the data.

B.2 Parameter Recovery

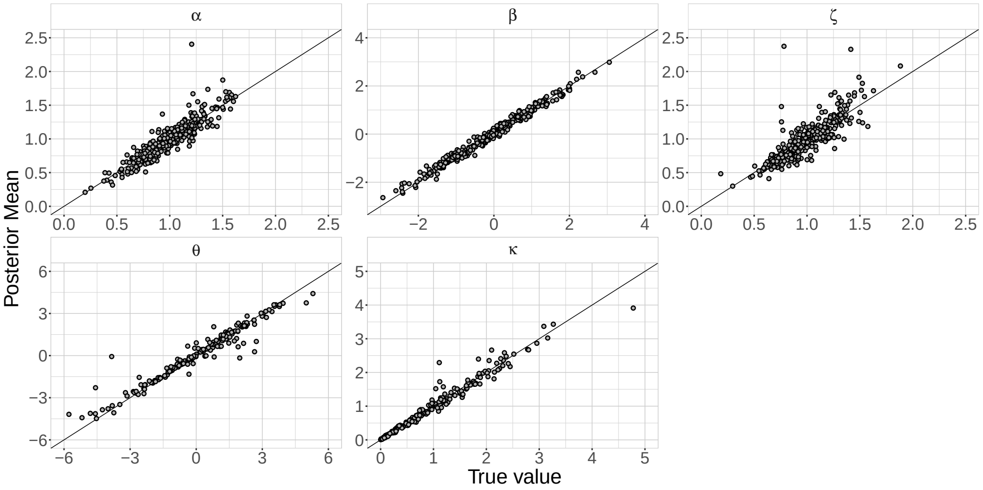


Figure B.3: Parameter recovery for the Latent Truth Rater model displayed in Figure 3.1. The data set consisted of 1 patient, 200 items, and 300 raters. Items had 5 possible outcomes.

C
Appendix of Chapter 4

Table C.1: Overview of the 22 IFTE Items, the factor on which they load, and the origin of the question. The first item is the that is treated as the 23$^{\text{rd}}$ item here. In the third column, "Prop." is short for "Proposed by clinicians". Adapted from Schuringa et al. (2014).

Item description	Factor	Origin
Has the patient changed in this last period?	-	-
Does the patient show problem insight?	Protective behaviors	HKT-R
Does the patient have psychotic symptoms?	Problematic behavior	HKT-R
Does the patient use any drugs or alcohol?	Problematic behavior	HKT-R
Does the patient show impulsive behavior?	Problematic behavior	HKT-R
Does the patient show antisocial behavior?	Problematic behavior	HKT-R
Does the patient show hostile behavior?	Problematic behavior	HKT-R
Does the patient show sufficient common social skills?	Resocialization skills	HKT-R
Does the patient show sufficient skills to take care of oneself?	Resocialization skills	HKT-R
Does the patient cooperate with your treatment?	Protective behaviors	HKT-R
Does the patient admit and take responsibility for the crime(s)?	Protective behaviors	HKT-R
Does the patient show adequate coping skills?	Protective behaviors	HKT-R
Does the patient comply with the rules and conditions of the center and/or the treatment?	Problematic behavior	HKT-R
Does the patient show sufficient labor skills?	Resocialization skills	HKT-R
Does the patient have antisocial associates?	Problematic behavior	HKT-R
Does the patient have balanced daytime activities?	Resocialization skills	HKT-R
Does the patient show sufficient financial skills?	Resocialization skills	Prop.
Does the patient use his medication in a consistent and adequate manner?	Protective behaviors	Prop.
Does the patient show sexual deviant behavior?	Problematic behavior	Prop.
Does the patient show manipulative behavior?	Problematic behavior	Prop.
Does the patient show skills to prevent drug and alcohol use?	Protective behaviors	ASP
Does the patient show skills to prevent physical aggressive behavior?	Protective behaviors	ASP
Does the patient show skills to prevent sexual deviant behavior?	Protective behaviors	ASP

D
Appendix of Chapter 5

D.1 Posterior Distribution for Effect Size under the Spike-and-Slab Model

The main text featured a paired samples t-test, both for the example and for the demonstration of regularities regarding the prior probability of the spike and the prior width of the slab. In this online Appendix we detail the prior distributions for this t-test and explain how the spike-and-slab shrinkage is related to Bayes factors. More generally, we show to derive the posterior distribution for effect size δ under the spike-and-slab model. We first derive the results for the slab and spike individually and combine them afterwards.

Following Rouder et al. (2018), we assume that the observed differences between the paired samples, denoted Z_i, are normally distributed with unknown mean δ and a variance of 1. As prior distribution for δ we use a normal distribution with mean 0 and variance σ^2. This implies $Z_i \sim \mathcal{N}(\delta, 1)$ for the data and $\delta \sim \mathcal{N}\left(0, \sigma^2\right)$ for the prior. The posterior distribution for δ is obtained through Bayes' theorem:

$$\overbrace{p\left(\delta \mid \boldsymbol{Z}\right)}^{\substack{\text{Posterior}\\ \text{distribution}}} = \overbrace{p\left(\delta\right)}^{\substack{\text{Prior}\\ \text{distribution}}} \times \frac{\overbrace{p(\boldsymbol{Z} \mid \delta)}^{\text{Likelihood}}}{\underbrace{p\left(\boldsymbol{Z}\right)}_{\substack{\text{Marginal}\\ \text{Likelihood}}}}.$$

The likelihood is given by:

$$\begin{aligned} p(\boldsymbol{Z} \mid \delta, \text{slab}) &= \prod_{i=1}^{N} \mathcal{N}\left(Z_i \mid \delta, 1\right) \\ &= (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s_{\boldsymbol{Z}}^2 + \delta^2 - 2\bar{\boldsymbol{Z}}\delta\right)\right), \end{aligned}$$

where $\bar{\boldsymbol{Z}}$ and $s_{\boldsymbol{Z}}^2$ are the sample mean and sample variance of Z_i respectively. Next, we compute the marginal likelihood by integrating out the likelihood times prior with respect to δ:

$$\begin{aligned}
p\left(\boldsymbol{Z} \mid \text{slab}\right) &= \int_{-\infty}^{\infty} p(\boldsymbol{Z} \mid \delta)\, p\left(\boldsymbol{Z}\right) \ \mathrm{d}\delta \\
&= (2\pi)^{-\frac{N+1}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}}\right)\right), \\
&\times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{\boldsymbol{Z}}}{\sigma^2}\right)\right) \ \mathrm{d}\delta.
\end{aligned}$$

Here we may recognize a Gaussian integral and use the following identity:

$$\int_{-\infty}^{\infty} \exp\left(-ax^2 + bx + c\right) \ \mathrm{d}x = \sqrt{\frac{\pi}{a}} \exp\left(\frac{b^2}{4a} + c\right).$$

Filling in the identity and simplifying yields:

$$p\left(\boldsymbol{Z} \mid \text{slab}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}}\right)\right) \frac{\exp\left(\frac{N^2\bar{\boldsymbol{Z}}^2}{2\left(N+\frac{1}{\sigma^2}\right)}\right)}{\sqrt{N + \frac{1}{\sigma^2}}}.$$

Next, we can obtain an expression for the posterior distribution. However, often it suffices to write out the expression for the likelihood times prior and then identify the result as a known distribution. This is particularly common in Gibbs sampling where one is interested in the conditional posterior distributions. We also do this here, as it reduces inference about the posterior distribution (e.g., what is the mean or variance) to inference about a known distribution, in this case a normal distribution:

$$\begin{aligned}
p\left(\delta \mid \boldsymbol{Z}, \text{slab}\right) &\propto (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}}\right)\right) \exp\left(-\frac{N}{2}\left(\delta^2 - 2\bar{\boldsymbol{Z}}\delta\right)\right) \\
&\times (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}\delta^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\delta^2\left(N + \frac{1}{\sigma^2}\right) - \delta\frac{2N\bar{\boldsymbol{Z}}}{\sigma^2}\right)\right).
\end{aligned}$$

We recognize a normal distribution with variance $\sigma_1^2 = \frac{1}{N+\frac{1}{\sigma^2}}$ and mean $\mu_1 = N\bar{\boldsymbol{Z}}\sigma_1^2$. Thus we have $p\left(\delta \mid \boldsymbol{Z}\right) \propto \mathcal{N}\left(\mu_1, \sigma_1^2\right)$.

Next we compute the same for the spike. The spike states that $Z_i \sim \mathcal{N}\left(0, 1\right)$ and contains no parameters to estimate. Thus there are no prior distributions to specify and all that needs to be computed is the marginal likelihood:

$$p\left(\boldsymbol{Z} \mid \text{spike}\right) = (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{N}{2}\left(\bar{\boldsymbol{Z}} + s^2_{\boldsymbol{Z}}\right)\right).$$

Using both marginal likelihoods we can now obtain the Bayes factor in favor of the spike:

$$\mathrm{BF}_{01} = \frac{p\left(\boldsymbol{Z} \mid \text{spike}\right)}{p\left(\boldsymbol{Z} \mid \text{slab}\right)} = \frac{\sqrt{N + \frac{1}{\sigma^2}}}{\exp\left(\frac{N^2 \bar{\boldsymbol{Z}}^2}{2\left(N + \frac{1}{\sigma^2}\right)}\right)}.$$

The posterior probability of the slab then equals:

$$p\text{slab} \mid \boldsymbol{Z} = \frac{p\text{spike}}{p\text{spike} + (1 - p\text{spike})\mathrm{BF}_{01}},$$

and the posterior probability of the spike is the complement. It then follows that the cumulative distribution function for the spike-and-slab posterior is given by:

$$P(\delta \leq x \mid \boldsymbol{Z}) = \begin{cases} p\text{slab} \mid \boldsymbol{Z}\Phi(x; \mu_1, \sigma_1) & \text{if } x < 0, \\ p\text{spike} \mid \boldsymbol{Z} + p\text{slab} \mid \boldsymbol{Z}\Phi(x; \mu_1, \sigma_1) & \text{if } x \geq 0, \end{cases}$$

where $\Phi(x; \mu_1, \sigma_1)$ is the cumulative normal distribution. Due to the discontinuity at $x = 0$ there is no useful closed form expression for the posterior density. Nevertheless, the posterior mean of the spike-and-slab model is available in closed form. Using the law of total probability, we have:

$$p(\delta \mid \boldsymbol{Z}) = p\text{spike} \mid \boldsymbol{Z}p(\delta \mid \text{spike}, \boldsymbol{Z}) + p\text{slab} \mid \boldsymbol{Z}p(\delta \mid \text{slab}, \boldsymbol{Z}).$$

Computing the mean of left hand side yields:

$$\begin{aligned} \int_{-\infty}^{\infty} \delta\, p(\delta \mid \boldsymbol{Z}) \;\mathrm{d}\delta &= p\text{spike} \mid \boldsymbol{Z} \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{spike}, \boldsymbol{Z}) \;\mathrm{d}\delta, \\ &\quad + p\text{slab} \mid \boldsymbol{Z} \int_{-\infty}^{\infty} \delta\, p(\delta \mid \text{slab}, \boldsymbol{Z}) \;\mathrm{d}\delta, \\ &= 0 + p\text{slab} \mid \boldsymbol{Z} \left(\mu_\delta \mid \text{slab}, \boldsymbol{Z}\right). \end{aligned}$$

Here $(\mu_\delta \mid \text{slab}, \boldsymbol{Z})$ is the posterior mean of effect size under the slab. In a similar fashion, other statistics may be obtained. However, it is also possible to draw samples from marginal posterior distribution. To obtain a sample s, first draw u from a uniform distribution on [0, 1]. If $u < p\text{slab} \mid \boldsymbol{Z}$ draw s from $p(\delta \mid \text{slab}, \boldsymbol{Z})$, otherwise s is zero. This approach is often used when the integrals become too unwieldy to compute analytically. For example, the R package BAS uses this procedure to compute credible intervals (Clyde et al., 2011a).

E
Appendix of Chapter 6

E.1 Jeffreys's Bayes Factor for the Agreement of Two Standard Errors

Our work was inspired by Jeffreys (1939, pp. 222-224), who developed a test for the "agreement of two standard errors". Specifically, let σ_1 and σ_2 be the standard errors for the two groups, respectively. Jeffreys estimates the standard errors by the expectation of the respective sum of squares, $(n_1-1)\sigma_1^2$ and $(n_2-1)\sigma_2^2$, where n_1 and n_2 are the respective sample sizes. Under the null hypothesis, the expectations are pooled such that $\lambda = (n_1+n_2-2)\sigma_1^2$, where $\sigma_1^2 = \sigma_2^2$. Under the alternative hypothesis, we have $\lambda = (n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2$, which can be written as a mixture such that $(n_1-1)\sigma_1^2 = \vartheta\lambda$ and $(n_2-1)\sigma_2^2 = (1-\vartheta)\lambda$. Because λ is common to both models, we can assign it an improper prior and integrate it out. The test-relevant parameter is $\vartheta \in [0,1]$, which Jeffreys assigns a uniform prior. After Laplace-approximating the integral under the alternative, Jeffreys arrives at the (approximate) Bayes factor:

$$\mathrm{BF}_{01}^{J} = \frac{(N-2)^{3/2}}{2\sqrt{\pi(n_1-1)(n_2-1)}} \exp\left(2\frac{n_2-n_1}{N-2}z - \frac{(n_1-1)(n_2-1)}{N-2}z^2\right) \quad , \tag{E.1.1}$$

where $N = n_1+n_2$ and $z = \log\left(\frac{s_1}{s_2}\right)$, and where s_1 and s_2 are the sample standard deviations.

As a side note, we first attempted a parameterization that, unbeknownst to us, Jeffreys substituted for his 1939 averaging idea in the third edition of the *Theory of Probability* (Jeffreys, 1961): $\sigma_1^2 = \sigma_2^2 e^{\xi}$. We abandoned this idea because we could not generalize it to $K > 2$ groups and instead adopted Jeffreys's original averaging idea.

Figure E.1 shows that our Bayes factor with $u = 1$ matches Jeffreys's 1939 Bayes factor very closely, as is expected from the uniform prior on ϑ. The error is due to his approximate solution. For completeness, we also show Jeffreys's 1961 Bayes factor, which is not limit consistent. It strikes us as a curiosity that Jeffreys would develop a test for the standard error instead of the population

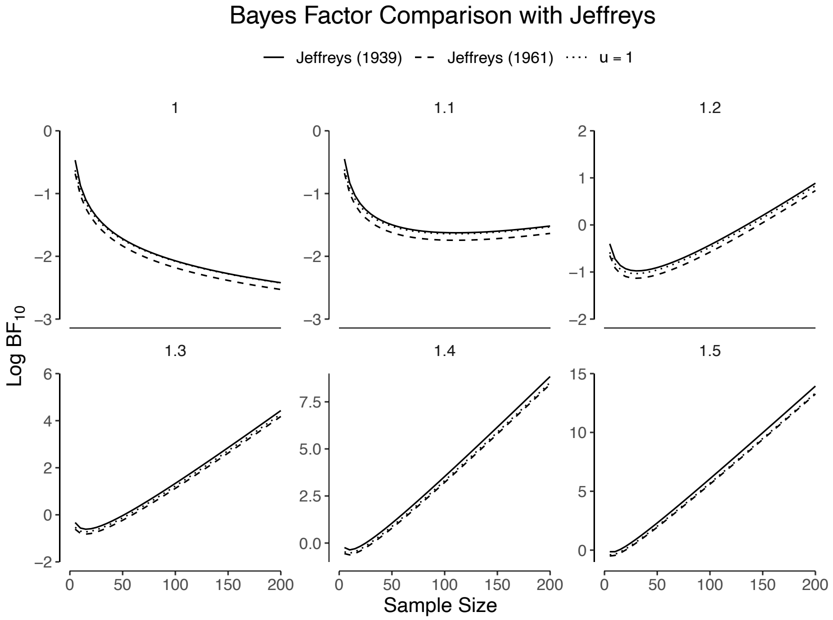


Figure E.1: Comparison of the Bayes factor proposed by Jeffreys (1939) and our Bayes factor with $u = 1$ for $K = 2$ groups as a function of the sample size and the effect size $\phi = \{1, 1.1, 1.2, 1.3, 1.4, 1.5\}$.

variance. Since the standard error decreases with the (square root of) the sample size, applying Jeffreys's test to data of unequal group sizes confounds the result (if we were to take his test as a test concerning equality of variances). Formally, both Bayes factors Jeffreys derived are not limit consistent because if we gather infinite data for only one group, the Bayes factor in favor of $\mathcal{H}_1$ will go to infinity instead of converging to a bound (Ly, 2018, ch. 6). For our Bayes factor, we adopt Jeffreys's averaging idea to parameterize the problem, but we focus on the population precisions instead of the standard errors.

E.2 Derivation of the proposed Bayes factor

E.2.1 Integrating out the nuisance parameters

Let $Y_{ji} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_j, \tau_j^{-1})$, where $i = 1, 2, \ldots, n_j$ and $j \in [K]$. For both the null and the alternative models we integrate the nuisance parameters μ_js out with respect to the right Haar priors $\mu_j \propto 1$. This implies that for the observations $y^{\{j\}}$ from the jth group consisting of n_j observations the likelihood function is

$$f(y^{\{j\}} \mid \tau_j) := \int f(y^{\{j\}} \mid \mu_j, \tau_j)\pi(\mu_j)\mathrm{d}\mu_j, \tag{E.2.1}$$

$$= (2\pi)^{-\frac{n_j}{2}} \tau_j^{\frac{n_j}{2}} \exp(-\tfrac{1}{2}\nu_j s_j^2 \tau_j) \int \exp(-\tfrac{n}{2}\tau_j(\bar{y}_j - \mu_j)^2)\mathrm{d}\mu_j, \tag{E.2.2}$$

$$= (2\pi)^{-\frac{\nu_j}{2}} n_j^{-\frac{1}{2}} \tau_j^{\frac{\nu_j}{2}} \exp(-\tfrac{1}{2}\nu_j s_j^2 \tau_j). \tag{E.2.3}$$

For data from the K samples combined, i.e., $y^{[K]}$, and the parametrisation $\tau_j = \vartheta_j \bar{\boldsymbol{\tau}} K$ this yields

$$f(y^{[K]} \mid \vec{\vartheta}, \bar{\boldsymbol{\tau}}) = (2^{-1}K)^{-\frac{\boldsymbol{\nu}_+}{2}} C(n) \Big[\prod_{j=1}^{K} \vartheta_j^{\frac{\nu_j}{2}} \Big] \bar{\boldsymbol{\tau}}^{\frac{\boldsymbol{\nu}_+}{2}} \exp\big(-2^{-1}K\bar{\boldsymbol{\tau}} \sum_{j=1}^{K} \vartheta_j \nu_j s_j^2 \big), \tag{E.2.4}$$

where $C(n) = (2\pi)^{-\boldsymbol{\nu}_+/2}(n_1 \ldots n_K)^{1/2}$ and $\boldsymbol{\nu}_+ = \sum_{j=1}^{K} \nu_j$. A natural prior on the nuisance parameter $\bar{\boldsymbol{\tau}}$ is $\pi(\bar{\boldsymbol{\tau}}) \propto \bar{\boldsymbol{\tau}}^{-1}$ and a standard gamma integral leads to the marginalized likelihood

$$\tilde{h}(y^{[K]} \mid \vec{\vartheta}) = \int f(y^{[K]} \mid \vec{\vartheta}, \bar{\boldsymbol{\tau}})\pi(\bar{\boldsymbol{\tau}})\mathrm{d}\bar{\boldsymbol{\tau}} = C(n)\Gamma\big(\frac{\boldsymbol{\nu}_+}{2}\big) \Big[\prod_{j=1}^{K} \vartheta_j^{\frac{\nu_j}{2}} \Big] \Big(\sum_{j=1}^{K} \vartheta_j \nu_j s_j^2 \Big)^{-\frac{\boldsymbol{\nu}_+}{2}}. \tag{E.2.5}$$

Since $\vartheta_j > 0$ and $\sum_{j=1}^{K} \vartheta_j = 1$ the vector $\boldsymbol{\vartheta} := (\vartheta_1, \ldots, \vartheta_K)$ can be fully described by $K-1$ free parameters. Any ϑ_j can be singled out in the following, but for concreteness, we do so for the Kth one. To rewrite the marginalized likelihood $\tilde{h}(y^{[K]} \mid \vec{\vartheta})$ in terms of the $K-1$ proportions ϑ, note that

$$\sum_{j=1}^{K} \vartheta_j \nu_j s_j^2 = \vartheta_1 \nu_1 s_1^2 + \vartheta_2 \nu_2 s_2^2 + \ldots + \vartheta_{K-1} \nu_{K-1} s_{K-1}^2 + \big(1 - \sum_{j=1}^{K-1} \vartheta_j\big) \nu_K s_K^2 \tag{E.2.6}$$

$$= \nu_K s_K^2 - \sum_{j=1}^{K-1} [\nu_K s_K^2 - \nu_j s_j^2]\vartheta_j, \tag{E.2.7}$$

which implies that

$$\Big(\sum_{j=1}^{K} \vartheta_j \nu_j s_j^2 \Big)^{-\frac{\boldsymbol{\nu}_+}{2}} = (\nu_K s_K^2)^{-\frac{\boldsymbol{\nu}_+}{2}} \Big(1 - \sum_{j=1}^{K-1} [1 - \tfrac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j \Big)^{-\frac{\boldsymbol{\nu}_+}{2}}. \tag{E.2.8}$$

This leads to

$$\tilde{h}(y^{[K]} \mid \vec{\vartheta}) = C(n)\Gamma\big(\frac{\boldsymbol{\nu}_+}{2}\big)(\nu_K s_K^2)^{-\frac{\boldsymbol{\nu}_+}{2}} \Big[\prod_{j=1}^{K} \vartheta_j^{\frac{\nu_j}{2}}\Big]\Big(1 - \sum_{j=1}^{K-1}[1 - \tfrac{\nu_j s_j^2}{\nu_K s_K^2})]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}, \tag{E.2.9}$$

which will be used to derive desiderata on the prior on the test relevant parameters. To highlight the fact that $\vec{\vartheta}$ is effectively $K-1$ dimensional, we can replace $\Big[\prod_{j=1}^{K} \vartheta_j^{\frac{\nu_j}{2}}\Big] = \Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1-\vec{\vartheta}_+)^{\frac{\nu_K}{2}}$, where $\vec{\vartheta}_+ := \sum_{j=1}^{K-1}\vartheta_j$.

E.2.2 Deriving the proposed Bayes factors

The marginalized likelihood fully specifies the marginal likelihood of the null, as the plugin $\vartheta_j = 1/K$ yields

$$p(y^{[K]} \mid \mathcal{M}_0) = C(n)\Gamma\big(\frac{\boldsymbol{\nu}_+}{2}\big)(\nu_K s_K^2)^{-\frac{\boldsymbol{\nu}_+}{2}} \Big(1 + \sum_{j=1}^{K-1} \frac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{-\frac{\boldsymbol{\nu}_+}{2}}. \tag{E.2.10}$$

We let $h(y^{[K]} \mid \vec{\vartheta}) = \frac{\tilde{h}(y^{[K]} \mid \vec{\vartheta})}{\tilde{h}(y^{[K]} \mid \vec{\vartheta} = \frac{1}{K})}$ be the reduced likelihood, see Eq. (6.1.7), and the Bayes factor is then

$$\mathrm{BF}_{10}(y^{[K]}) = \Big(1 + \sum_{j=1}^{K-1} \frac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} \tag{E.2.11}$$

$$\times \int \Big[\prod_{j=1}^{K-1} \vartheta_j^{\frac{\nu_j}{2}}\Big](1-\vec{\vartheta}_+)^{\frac{\nu_K}{2}} \Big(1 - \sum_{j=1}^{K-1}[1 - \tfrac{\nu_j s_j^2}{\nu_K s_K^2}]\vartheta_j\Big)^{-\frac{\boldsymbol{\nu}_+}{2}} \pi_1(\vec{\vartheta})\mathrm{d}\vec{\vartheta}, \tag{E.2.12}$$

where $\vec{\vartheta} \in \mathbb{R}^{K-1}$, and the integral is over the $K-1$ simplex. A natural prior for $\vec{\vartheta}$ would be a Dirichlet prior with hyperparameters $\boldsymbol{u}$, where $\boldsymbol{u} = (u_1, \ldots, u_{K-1}, u_K)$ with non-negative components. For $\nu_j \geq 1$ for all $j \in [K]$ and by definition of the multivariate integral representation of the type D Lauricella function of $K-1$ variables (Lauricella, 1893), this Bayes factor is analytic and given by

$$\mathrm{BF}_{10}(y^{[K]}) = \frac{\mathcal{B}(\frac{\vec{\nu}}{2}+\vec{u})}{\mathcal{B}(\vec{u})}\Big(1 + \sum_{j=1}^{K-1} \frac{\nu_j s_j^2}{\nu_K s_K^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} F_D\Big(\tfrac{\boldsymbol{\nu}_+}{2}\,;\, \tfrac{\vec{\nu}}{2} + \vec{u};\, \tfrac{\boldsymbol{\nu}_+}{2} + u_+\,;\, \vec{1} - \tfrac{\overrightarrow{\nu s^2}}{\nu_K s_K^2}\Big) \tag{E.2.13}$$

where $\mathcal{B}(\vec{u}) = \frac{\Gamma(u_1)\cdots\Gamma(u_K)}{\Gamma(u_+)}$ is the multivariate beta function, $\vec{1} = (1, \ldots, 1) \in \mathbb{R}^{K-1}$ and where $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1}s_{K-1}^2)$ is the $K-1$ vector of sums of squares.

E.2.3 Default approach to non-overlapping Bayes factors

Note that non-overlapping hypotheses can also be expressed in terms of the location parameter $\delta = -\log(\frac{\vartheta}{1-\vartheta}) \in \mathbb{R}$, which transforms the point null hypothesis $\mathcal{H}_0 : \vartheta = 1/2$ to $\mathcal{H}_0 : \delta = 0$ yielding a comparison between

$$\breve{\mathcal{H}}_0 : |\delta| < \epsilon \text{ and } \breve{\mathcal{H}}_1 : |\delta| > \epsilon, \tag{E.2.14}$$

where ϵ defines the half width of the null-region. Berger and Delampady (1987) showed that for the location problem $\bar{X} \sim \mathcal{N}(\delta, \sigma^2/n)$ with a unimodal and symmetric prior, and $\epsilon \leq \frac{\sigma}{2\sqrt{n}}$, the standard (point null) Bayes factor characterizes the behavior of the null-region Bayes factor comparing $\breve{\mathcal{H}}_1$ to $\breve{\mathcal{H}}_0$ with the priors truncated accordingly.

Note that the prior $\vartheta \sim \text{Beta}(u, u)$ underlying Eq. (6.3.1) induces a type III generalized logistic distribution on δ with density

$$\pi(\delta) = \tfrac{1}{\mathcal{B}(u,u)} e^{-\delta u}(1 + e^{-\delta})^{-2u}. \tag{E.2.15}$$

This prior is unimodal and symmetric around $\mathcal{H}_0 : \delta = 0$. In terms of δ the marginalized likelihood $\tilde{h}(y^{[K]} \mid \vartheta)$, see appendix Eq. (E.2.8), is

$$\tilde{h}(y^{[K]} \mid \delta) \propto \exp(-ng(\delta)), \text{ where } g(\delta) \approx \tfrac{c}{2}\delta + \tfrac{1+c}{2}\log(1 + \tfrac{s_1^2}{s_2^2} c e^{-\delta}), \tag{E.2.16}$$

whenever $n_1 = cn$ and $n_2 = n$. Sufficiently large n combined with a Taylor expansion of $g(\delta)$ at its maximum point, that is, at $\hat{\delta} = \log(\frac{s_1^2}{s_2^2})$, yields the approximation

$$\tilde{h}(y^{[K]} \mid \delta) \propto \exp\Big(-\tfrac{nc}{4(1+c)}\big(\delta - \log(s_1^2/s_2^2)\big)^2\Big). \tag{E.2.17}$$

Hence, one way to take a null interval is by setting $\epsilon \leq \frac{(1+c)}{\sqrt{nc}}$. The resulting null-region Bayes factor will then behave similarly to Eq. (6.3.1).

E.3 Properties of the proposed Bayes factor

E.3.1 Labelling Invariant

Proof of labelling invariance, Theorem 6.2.1. The goal is to show that the integral of the reduced likelihood times prior remains the same after applying the permutation ϱ that swaps the labels K for an arbitrary $i \in [K-1]$. For this integral to remain the same, it suffices to show that the reduced likelihood

$h(\boldsymbol{s^2} \mid \vec{\vartheta})$ and its permuted version

$$h(\varrho(\boldsymbol{s^2}) \mid \vec{\vartheta}) = \Big(1 + \tfrac{\nu_K s_K^2}{\nu_i s_i^2} + \sum_{j \in [K-1] \setminus \{i\}} \tfrac{\nu_j s_j^2}{\nu_i s_i^2}\Big)^{\frac{\boldsymbol{\nu}_+}{2}} \Big[\prod_{j \in [K-1] \setminus \{i\}} \vartheta_j^{\frac{\nu_j}{2}} \Big] \tag{E.3.1}$$

$$\times \vartheta_i^{\frac{\nu_K}{2}} (1 - \vec{\vartheta}_+)^{\frac{\nu_i}{2}} \Big(1 - \vec{\theta}_+ + \tfrac{\nu_K s_K^2}{\nu_i s_i^2} \vartheta_i + \sum_{j \in [K-1] \setminus \{i\}} \tfrac{\nu_j s_j^2}{\nu_i s_i^2} \vartheta_j \Big)^{-\frac{\boldsymbol{\nu}_+}{2}}, \tag{E.3.2}$$

are conditionally symmetric. This means that as a function of ϑ_i with all other coordinates fixed, i.e., ϑ_j for $j \in [K-1] \setminus \{i\}$, the reduced likelihood and its permuted version are symmetric around $\breve{\vartheta}_{-i} := \frac{1}{2}\big(1 - \sum_{j \in [K-1] \setminus \{i\}} \vartheta_j\big)$.

This can be shown by studying the functions $g(x)$ and $g_\varrho(-x)$, where $g(x)$ is the composition of $x \mapsto \vartheta_i = \breve{\vartheta}_{-i} + x$ and $\vartheta_i \mapsto h(\boldsymbol{s^2} \mid \vec{\vartheta})$, whereas $g_\varrho(-x)$ is the composition of $x \mapsto \vartheta_i = \breve{\vartheta}_{-i} - x$ and $\vartheta_i \mapsto h(\varrho(\boldsymbol{s^2}) \mid \vec{\vartheta})$. A straightforward, but tedious computation then shows that $g(x) = g_\varrho(-x)$ for all $x \in (0, \breve{\vartheta}_{-i})$. For the Bayes factor to be labelling invariant, we thus require the prior to be symmetric in the similar fashion. For the Dirichlet prior this implies $u_i = u_K$, and for this to hold for all pairs of permutations, we require $u_j = u$ for all $j \in [K]$. □

E.3.2 Predictive Matching

Proof of predictive matching, Theorem 6.2.2. Case (a) with $n_1 = \ldots = n_K = 1$ implies that $\nu_1 s_1^2 = \ldots = \nu_K s_K^2 = 0$ regardless of the data, which implies that the likelihood of the data Eq. (E.2.4) is identical to the constant function 1, thus, independent of $\bar{\boldsymbol{\tau}}$ and $\vec{\vartheta}$. Viewing the prior $\bar{\boldsymbol{\tau}} \propto \bar{\boldsymbol{\tau}}^{-1}$ on the nuisance parameter that appears in both the numerator and the denominator of the Bayes factor as a limit of $\bar{\boldsymbol{\tau}} \sim \Gamma(u, u)$ with $u \downarrow 0$ shows that without loss of generality we can set the Bayes factor to 1, whenever $\pi_1(\vec{\vartheta})$ is proper.

For case (b) and without loss of generality we consider the case with $\nu_K = 1$ and $\nu_j = 0$ for all $j \in [K-1]$. The reduced likelihood $h(\boldsymbol{s^2} \mid \vec{\vartheta})$ is then actually independent of s_K^2, as we then get

$$\mathrm{BF}_{10}(\boldsymbol{s^2}) = \frac{\int (s_K^2)^{-\frac{1}{2}} (1 - \vec{\vartheta}_+)^{\frac{1}{2}} (1 - \vec{\vartheta}_+)^{-\frac{1}{2}} \pi_1(\vec{\vartheta}) \mathrm{d}\vartheta}{(s_K^2)^{-\frac{1}{2}} (\frac{1}{K})^{\frac{1}{2}} (1 - \frac{K-1}{K})^{-\frac{1}{2}}} = \int \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}. \tag{E.3.3}$$

Thus, for all data sets $\boldsymbol{s^2}$ of insufficient size $\mathrm{BF}_{10}(\boldsymbol{s^2}) = 1$ whenever $\pi_1(\vec{\vartheta})$ is proper. □

E.3.3 Information Consistency

Proof of information consistency, Theorem 6.2.3. Assuming labelling invariance we can let the s_K^2 with fixed n_K grow without loss of generality. For fixed $\boldsymbol{n}$ the order of integral and limit can be interchanged and reveals that

$$\lim_{s_K^2 \to \infty} \tag{E.3.4}$$

The integrand becomes unbounded whenever $u_K \leq \frac{\nu_+ - \nu_K}{2}$. Recall that the minimal sample size has only two groups with two observations, say, $\nu_1 = 1$ and $\nu_K = 1$. The requirement that $\lim_{s_K^2 \to \infty} \mathrm{BF}_{10}(\boldsymbol{s^2})$ should already diverge at the minimal sample sizes implies that $u_K \leq 1/2$. By symmetry we require this for all u_j for $j \in [K]$. □

E.3.4 Model selection consistency

For model selection consistency we note that the Bayes factor depends on the data via the statistic $\vec{W} = (W_1, \ldots, W_{K-1})$ with

$$W_j := \frac{\nu_j s_j^2}{\nu_K s_K^2} = \frac{\sigma_j^2 \nu_j}{\sigma_K^2 \nu_K} \frac{\left(\sum_{i=1}^{n_j} \frac{(Y_{ji} - \bar{Y}_j)^2}{\sigma_j^2} \right) / \nu_j}{\left(\sum_{i=1}^{n_K} \frac{(Y_{Ki} - \bar{Y}_K)^2}{\sigma_K^2} \right) / \nu_K} =: \frac{\sigma_j^2 \nu_j}{\sigma_K^2 \nu_K} X_j, \text{ for } j \in [K-1], \tag{E.3.5}$$

where $X_j \sim F(\nu_j, \nu_K)$ is an F-distributed random variable with degrees of freedom ν_j and ν_K by virtue of the data being normally distributed.

Letting $n_j := c_j n$ for $c_j > 0$, $j \in [K]$, thus, $c_K = 1$, and $\sigma_j^2 := \gamma_j \sigma_K^2$ where $\gamma_j > 0$ for $j \in [K]$, thus, $\gamma_K = 1$, note that $W_j \approx c_j \gamma_j X_j$ for n large. Observe that since X_j is F-distributed we know that

$$E(X_j) = \frac{n}{n-2} = 1 + \mathcal{O}(1/n) \text{ and } \mathrm{Var}(X_j) = \frac{2n^2((1+c_j)n - 2)}{c_j n (n-2)^2 (n-4)} = \mathcal{O}(1/n). \tag{E.3.6}$$

Hence, Chebyshev's inequality can be applied to show that $X_j - 1 = \mathcal{O}_P(n^{-1/2})$. The intuition to use the continuous mapping theorem and the replacement $\vec{X} = \vec{1} \in \mathbb{R}^{K-1}$ in BF_{10} forms the basis of the proof of Theorem 6.2.4. What needs taking care of is the dependence of the Bayes factor on n.

Proof of model selection consistency, Theorem 6.2.4. The proof relies on a Taylor approximation that holds with high probability and the subsequent asymptotic analysis of the Taylor terms. Key to this analysis is the large sample behavior of gamma functions. What is remarkable is that under the null the exponential growing terms cancelled out perfectly in all Taylor terms.

Notation for partial derivatives For the Taylor terms, we express the Bayes factor as follows $\mathrm{BF}_{10}(\boldsymbol{s}^2, n) = \frac{\mathcal{B}(\frac{n}{2}\boldsymbol{c}+\boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} b(\vec{X}) G_D(\vec{X})$, where with $\vec{Z} \in \mathbb{R}^{K-1}$, $Z_j = 1 - c_j \gamma_j X_j$, the $\vec{X}$-dependent functions are

$$b(\vec{X}) := (1 + \sum_{j=1}^{K-1} c_j \gamma_j X_j)^{\frac{\boldsymbol{c}_+}{2} n}, \tag{E.3.7}$$

$$G_D(\vec{X}) := F_D(\tfrac{\boldsymbol{c}_+}{2} n\,;\, \tfrac{n}{2}\vec{c} + \vec{u}\,;\, \tfrac{\boldsymbol{c}_+}{2} n + \boldsymbol{u}_+\,;\, \vec{Z}). \tag{E.3.8}$$

For the Taylor series we employ multi-index notation to describe Leibniz's product rule for partial derivatives. The idea is to identify a partial derivative to a $K-1$-dimensional vector of non-negative integers $\vec{m} \in \mathbb{N}_0^{K-1}$. Each m_j represents the multiplicity of partial derivative with respect to the variable x_j, thus, $\partial^{\vec{m}} b(\vec{X}) := \frac{\partial^{\vec{m}_+}}{\prod_{j=1}^{K-1} \partial x_j^{m_j}} b(\vec{X})$ and more specifically

$$\partial^{\vec{m}} b(\vec{X}) = (\tfrac{\boldsymbol{c}_+}{2} n)_{-\vec{m}_+} \Big(\prod_{j=1}^{K-1} (c_j \gamma_j)^{m_j} \Big) (1 + \sum_{j=1}^{K-1} c_j \gamma_j X_j)^{\frac{\boldsymbol{c}_+}{2} n - \vec{m}_+}, \tag{E.3.9}$$

where $(a)_{-l} := \Gamma(a+1)/\Gamma(a-l+1)$ denotes the falling factorial, e.g., $(a)_{-3} = a(a-1)(a-2)$ for $a \in \mathbb{N}$. It can be shown that $(a)_{-l} = (-1)^l (-a)_l$ and that $(a)_{-l}/l! = \binom{a}{l}$. Note that $b(\vec{X})$ also appears on the right-hand side. To simplify notation we write

$$\partial^{\vec{m}} b := \partial^{\vec{m}} b(\vec{X}) \Big|_{\vec{X}=\vec{1}} = \big(\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle\big)^{\frac{\boldsymbol{c}_+}{2} n} \frac{(\frac{\boldsymbol{c}_+}{2} n)_{-\vec{m}_+} \Big(\prod_{j=1}^{K-1} (c_j \gamma_j)^{m_j} \Big)}{\big(\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle\big)^{\vec{m}_+}}. \tag{E.3.10}$$

Note that the first order partial derivatives are described by the vectors $\vec{m} = \vec{e}_k$ for $k \in [K-1]$.

Similarly, let $\vec{l} \in \mathbb{N}_0^{K-1}$ with $\vec{m} \preceq \vec{l}$, that is, $0 \leq m_j \leq l_j$ for $j \in [K-1]$, then $\vec{r} = \vec{l} - \vec{m} \in \mathcal{N}_0^{K-1}$ can be thought of as the remaining multiplicities of $\vec{l}$ once the partial derivatives are taken with multiplicities $\vec{m}$. This vector notation combined with differentiation under the integral sign shows that

$$\partial^{\vec{r}} G_D(\vec{X}) := \frac{\partial^{\vec{r}_+}}{\prod_{j=1}^{K-1} \partial x_j^{r_j}} G_D(\vec{X}), \tag{E.3.11}$$

$$= (-\tfrac{\boldsymbol{c}_+}{2} n)_{-\vec{r}_+} \frac{\prod_{j=1}^{K-1} (\frac{c_j}{2} n + u_j)_{r_j}}{(\frac{\boldsymbol{c}_+}{2} n + \boldsymbol{u}_+)_{\vec{r}_+}} \Big(\prod_{j=1}^{K-1} (c_j \gamma_j)^{r_j} \Big) G_{D,\vec{r}}(\vec{X}), \tag{E.3.12}$$

where, formally by Eq. (E.3.19) below,

$$\frac{\prod_{j=1}^{K-1} (\frac{c_j}{2} n + u_j)_{r_j}}{(\frac{\boldsymbol{c}_+}{2} n + \boldsymbol{u}_+)_{\vec{r}_+}} = \frac{\prod_{j=1}^{K-1} c_j^{r_j}}{\boldsymbol{c}_+^{\vec{r}_+}} \big(1 + \mathcal{O}(n^{-1})\big), \tag{E.3.13}$$

and where

$$G_{D,\vec{r}}(\vec{X}) = F_D(\tfrac{\boldsymbol{c}_+}{2} + \vec{r}_+ \,;\, \tfrac{n}{2}\vec{c} + \vec{u} + \vec{r};\, \tfrac{\boldsymbol{c}}{2}n + \boldsymbol{u}_+ + \vec{r}_+ \,;\, \vec{Z}). \tag{E.3.14}$$

Observe that $G_D(\vec{X}) = G_{D,\vec{0}}(\vec{X})$.

With this notation the partial derivative of the Bayes factor accounting for multiplicities $\vec{l}$ is

$$\partial^{\vec{l}}\mathrm{BF}_{10}(\boldsymbol{s}^2, n) = \frac{\mathcal{B}(\frac{n}{2}\boldsymbol{c} + \boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} \left(\sum_{\vec{m} \preceq \vec{l}} \binom{\vec{l}}{\vec{m}} \partial^{\vec{m}} b(\vec{X}) \partial^{\vec{l}-\vec{m}} G(\vec{X}) \right), \tag{E.3.15}$$

where $\binom{\vec{l}}{\vec{m}} = \binom{l_1}{m_1} \cdots \binom{l_{K-1}}{m_{K-1}} = \prod_{j=1}^{K-1} \frac{l_j!}{(l_j - m_j)! m_j!}$ and where the sum is over all subvectors $\vec{m}$ of $\vec{l}$. For instance, with $\vec{l} = \vec{e}_k$ this means $\vec{m} = \vec{0}$ and $\vec{m} = \vec{e}_k$. Note that $\partial^{\vec{l}}\mathrm{BF}_{10}(\boldsymbol{s}^2, n)$ only describes one entry of the $\vec{l}_+$-dimensional array of the total derivative of $\mathrm{BF}_{10}(\boldsymbol{s}^2, n)$ of order $\vec{l}_+$.

Taylor approximation Because the samples variances of the X_js are of order $1/n$, Chebyshev's inequality in conjunction with a union bound can be used to show that for any ϵ there exists an N such that if $n > N$ the following Taylor approximation holds with chance at least $1 - \epsilon$

$$\mathrm{BF}_{10}(\boldsymbol{s}^2, n) \approx \frac{\mathcal{B}(\frac{n}{2}\boldsymbol{c} + \boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} \Big(\sum_{\vec{l} \in \mathbb{N}_0^{K-1}} \partial^{\vec{l}} \big[bG(\vec{X}) \big]_{\vec{X} = \vec{1}} \frac{Q^{\vec{l}}}{\vec{l}!} \Big), \tag{E.3.16}$$

where $\partial^{\vec{l}}\big[bG(\vec{X})\big]_{\vec{X}}$ equals the sum on the right-hand side of Eq. (E.3.15) evaluated at $\vec{X} = \vec{1}$, $\vec{Q} = (\vec{X} - \vec{1})$, $\frac{\vec{Q}^{\vec{l}}}{\vec{l}!} = \prod_{j=1}^{K-1} \frac{Q_j^{l_j}}{l_j!}$. Below we will show that for large n the Bayes factor behaves as

$$\mathrm{BF}_{10}(\boldsymbol{s}^2, n) \approx \breve{T}^{(0)} \sum_{\vec{l} \in \mathbb{N}_0^{K-1}} h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}) \frac{\vec{Q}^{\vec{l}}}{\vec{l}!}, \tag{E.3.17}$$

where under the null $h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) = \mathcal{O}(1)$ and under the alternative $h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) = \mathcal{O}(n^{\vec{l}_+})$, and where $\breve{T}^{(0)}$ is the zeroth order term of the Taylor approximation studied in the next paragraph.

The $T^{(0)}$ term The large sample behavior of the Bayes factor basically follows from gamma function asymptotics. The first object of interest is the deterministic term associated with $\vec{l} = \vec{0}$, i.e., the Bayes factor evaluated at $\vec{X} = \vec{1}$, but still dependent on the n term is

$$T^{(0)} := \mathrm{BF}_{10}(\boldsymbol{s}^2, n)\Big|_{\vec{X}=\vec{1}} = \frac{\mathcal{B}(\frac{n}{2}\boldsymbol{c}+\boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})}(\langle \boldsymbol{c}, \boldsymbol{\gamma}\rangle)^{\frac{\boldsymbol{c}_+}{2}n} G_D. \tag{E.3.18}$$

The large sample behavior of the beta function follows that of gamma functions. Laplace's method implies that for $v, b > 0$

$$\Gamma(vn+b) = \sqrt{2\pi}(vn)^{vn+b-\frac{1}{2}}e^{-vn}\big[1 + \tfrac{6b^2-6b+1}{12}(vn)^{-1} + \mathcal{O}(n^{-2})\big] \tag{E.3.19}$$

as $n \to \infty$. Hence,

$$\mathcal{B}(\tfrac{n}{2}\boldsymbol{c}+\boldsymbol{u}) = (4\pi)^{\frac{K-1}{2}} n^{\frac{1-K}{2}} \boldsymbol{c}_+^{\frac{1}{2}} \Big(\prod_{j=1}^{K-1}(c_j)^{-\frac{1}{2}}\Big) g(\boldsymbol{c}, \boldsymbol{u}, n)\big[1 + \mathcal{O}(n^{-1})\big], \tag{E.3.20}$$

where the exponential behavior is captured by

$$g(\boldsymbol{c}, \boldsymbol{u}, n) = (\boldsymbol{c}_+)^{-\frac{\boldsymbol{c}_+ n}{2} - \boldsymbol{u}_+} \prod_{j=1}^{K-1} (c_j)^{\frac{c_j n}{2} + u_j}. \tag{E.3.21}$$

Note that the product only goes up to $K-1$, since $c_K = 1$ by definition.

The hard part is to show consistency under the null. For this the exponential behavior of $g(\boldsymbol{c}, \boldsymbol{u}, n)$ needs to be cancelled by that of G_D, and we will show that it does so perfectly. To study the large n behavior of G_D, and more generally $G_{D,\vec{r}}$, we apply a Pfaff transform (Lauricella, 1893, p. 148) yielding

$$G_{D,\vec{r}} = \Big(\prod_{j=1}^{K-1} c_j\gamma_j^{-\frac{c_j}{2}n - u_j - r_j}\Big) F_D\Big(\boldsymbol{u}_+\,;\, \tfrac{n}{2}\vec{c} + \vec{u} + \vec{r};\, \tfrac{\boldsymbol{c}_+}{2}n + \boldsymbol{u}_+ + \vec{r}_+\,;\, \overrightarrow{\tfrac{c\gamma - 1}{c\gamma}}\Big) \tag{E.3.22}$$

where $\overrightarrow{\frac{c\gamma-1}{c\gamma}} \in \mathbb{R}^{K-1}$ with $(\overrightarrow{\frac{c\gamma-1}{c\gamma}})_j = \frac{c_j\gamma_j - 1}{c_j\gamma_j}$. This rewrite of $G_{D,\vec{r}}$ shows a cancellation of the $(c_j\gamma_j)^{r_j}$ terms in front of the $G_{D,\vec{r}}$ in Eq. (E.3.12). Note that in the Lauricella function in Eq. (E.3.22) the lower term and the upper terms of the second kind depend on n in a linear fashion. The n dependence in these terms balance out as $n \to \infty$ making the Lauricella function in Eq. (E.3.22) of order 1 as n grows. This is made rigorous by Lemma 1, which shows that the Lauricella function Eq. (E.3.22) converges to a (generalized) negative binomial series as $n \to \infty$. Thus,

$$G_{D,\vec{r}} \approx \breve{G}_{D,\vec{r}} = \Big(\prod_{j=1}^{K-1}(c_j\gamma_j)^{-\frac{c_j}{2}n - u_j - r_j}\Big)\Big(1 - \tfrac{1}{\boldsymbol{c}_+}\sum_{j=1}^{K-1}\tfrac{c_j\gamma_j - 1}{\gamma_j}\Big)^{-\boldsymbol{u}_+}, \tag{E.3.23}$$

for n large. For $T^{(0)}$ set $\vec{r} = \vec{0}$, which shows that for large n

$$T^{(0)} \approx \breve{T}^{(0)} := C_0(K, \boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{u}) n^{\frac{1-K}{2}} \big(\tfrac{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle}{\boldsymbol{c}_+}\big)^{\frac{\boldsymbol{c}_+}{2} n} \Big(\prod_{j=1}^{K-1} \gamma_j^{-\frac{c_j}{2} n} \Big), \tag{E.3.24}$$

where the n independent term $C_0(K, \boldsymbol{\gamma}, \boldsymbol{c}, \boldsymbol{u})$ is as asserted in Eq. (6.2.3). A plugin of the null hypothesis $\boldsymbol{\gamma} = \mathbf{1}$, thus, $\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle = \boldsymbol{c}_+$, in Eq. (E.3.24) shows that the exponentially growing terms are all equal to one, and therefore $T^{(0)} = \mathcal{O}(n^{\frac{1-K}{2}})$.

The $T_{\vec{e}_k}^{(1)}$ terms The analysis of the gradient is similar to that of $T^{(0)}$. It suffices to study the gradient coordinate wise. In particular,

$$T_{\vec{e}_k}^{(1)} := \frac{\mathcal{B}(\frac{n}{2}\boldsymbol{c} + \boldsymbol{u})}{\mathcal{B}(\boldsymbol{u})} (\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle)^{\frac{\boldsymbol{c}_+}{2} n} c_k \gamma_k \frac{\boldsymbol{c}_+}{2} n \Big[\tfrac{G_D}{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle} - \tfrac{c_k n + 2 u_k}{\boldsymbol{c}_+ n + 2 \boldsymbol{u}_+} G_{D, \vec{e}_k} \Big]. \tag{E.3.25}$$

The same operations as before, a Pfaff transform and Eq. (E.3.23), shows that

$$\breve{T}_{\vec{e}_k}^{(1)} = \breve{T}^{(0)} \Big(\tfrac{\boldsymbol{c}_+}{2} \big(\tfrac{c_k \gamma_k}{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle} - \tfrac{c_k}{\boldsymbol{c}_+} \big) n + \tfrac{c_k \boldsymbol{u}_+ - \boldsymbol{c}_+ u_k}{\boldsymbol{c}_+} + \mathcal{O}(n^{-1}) \Big), \tag{E.3.26}$$

as $n \to \infty$. Hence, under the alternative $h_{\vec{e}_k}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) := \breve{T}_{\vec{e}_k}^{(1)} / \breve{T}^{(0)} = \mathcal{O}(n)$ and accounting for the stochastic term $Q_k = (X_k - 1) = \mathcal{O}_P(n^{-1/2})$ leads to $\sum_{k=1}^{K-1} h_{\vec{e}_k}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) Q_k = \mathcal{O}_P(n^{1/2})$. On the other hand, under the null $h_{\vec{e}_k}(\boldsymbol{u}, \boldsymbol{c}, \mathbf{1}, n) = \breve{T}_{\vec{e}_k}^{(1)} / \breve{T}^{(0)} = \mathcal{O}(1)$, as then again $\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle = \boldsymbol{c}_+$ and $\big(\tfrac{c_k \gamma_k}{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle} - \tfrac{c_k}{\boldsymbol{c}_+} \big) = 0$, thus, a perfect cancellation of the $\mathcal{O}(n)$ term. Consequently, $\sum_{k=1}^{K-1} h_{\vec{e}_k}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) Q_k = \mathcal{O}_P(n^{-1/2})$.

Higher order terms The higher order terms exhibit the same behavior. Let $\vec{l} \in \mathbb{N}^{K-1}$, then for n large the partial derivative associated to $\vec{l}$ of the Bayes factor behaves as

$$\breve{T}_{\vec{l}}^{(\vec{l}_+)} = \sum_{\vec{m} \preceq \vec{l}} \binom{\vec{l}}{\vec{m}} \breve{T}^{(0)} (\tfrac{\boldsymbol{c}_+}{2} n)_{-\vec{m}_+} (-\tfrac{\boldsymbol{c}_+}{2} n)_{-(\vec{l}_+ - \vec{m}_+)} \frac{\prod_{j=1}^{K-1} (c_j \gamma_j)^{m_j}}{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle^{\vec{m}_+}} \frac{\prod_{j=1}^{K} c_j^{l_j - m_j}}{\boldsymbol{c}_+^{(\vec{l}_+ - \vec{m}_+)}} \big(1 + \mathcal{O}(n^{-1})\big).$$

Note that $(\frac{\boldsymbol{c}_+}{2} n)_{-\vec{m}_+} (-\frac{\boldsymbol{c}_+}{2} n)_{-(\vec{l}_+ - \vec{m}_+)}$ is a polynomial in n of order $\vec{l}_+$. Hence, $h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \boldsymbol{\gamma}, n) := \breve{T}_{\vec{l}}^{(\vec{l}_+)} / \breve{T}_0 = \mathcal{O}(n^{\vec{l}_+})$. We now show that under the null, the polynomial $(\frac{\boldsymbol{c}_+}{2} n)_{-\vec{m}_+} (-\frac{\boldsymbol{c}_+}{2} n)_{-(\vec{l}_+ - \vec{m}_+)}$ is zero and $h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \mathbf{1}, n) = \breve{T}_{\vec{l}}^{(\vec{l}_+)} / \breve{T}_0 = \mathcal{O}(1)$, where the constant term comes from the approximation of the ratio of Pochhammer symbols, i.e., Eq. (E.3.13), e.g., Eq. (E.3.26). To see that there is no n contribution under the null, we plugin $\boldsymbol{\gamma} = \mathbf{1}$ and rewrite the sum over $\vec{m} \preceq \vec{l}$ as a sum over $\vec{m}_+ = p$ for $p = 0, 1, \ldots, \vec{l}_+$ and a subsequent sum over all subvector $\vec{m}$ that sum to p, which yields

$$h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \mathbf{1}, n) = \frac{\prod_{j=1}^{K-1} c_j^{l_j}}{\boldsymbol{c}_+^{\vec{l}_+}} \sum_{p=0}^{\vec{l}_+} (\tfrac{\boldsymbol{c}_+}{2} n)_{-p} (-\tfrac{\boldsymbol{c}_+}{2} n)_{-p} \sum_{\substack{\vec{m} \preceq \vec{l} \\ \vec{m}_+ = p}} \binom{\vec{l}}{\vec{m}}. \tag{E.3.27}$$

Next we apply the Chu-Vandermonde identity twice, once over the sum on the right-hand side of the previous display and once after using the identity $(a)_{-l}/l! = \binom{a}{l}$, which leads to

$$h_{\vec{l}}(\boldsymbol{u}, \boldsymbol{c}, \mathbf{1}, n) = \frac{\prod_{j=1}^{K-1} c_j^{l_j}}{\boldsymbol{c}_+^{\vec{l}_+}} \sum_{p=0}^{\vec{l}_+} \binom{\vec{l}_+}{p} (\tfrac{\boldsymbol{c}_+}{2} n)_{-p} (-\tfrac{\boldsymbol{c}_+}{2} n)_{-(\vec{l}_+ - p)} \tag{E.3.28}$$

$$= \frac{\prod_{j=1}^{K-1} c_j^{l_j}}{\boldsymbol{c}_+^{\vec{l}_+}} \vec{l}_+! \sum_{p=0}^{\vec{l}_+} \binom{\frac{\boldsymbol{c}_+}{2} n}{p} \binom{-\frac{\boldsymbol{c}_+}{2} n}{\vec{l}_+ - p} = 0. \tag{E.3.29}$$

This shows that under the null, none of the Taylor terms lead to a growth in n.

The stochastic terms in the assertion both under the null and the alternative follow from the definition of the exponential series by rewriting the sum of the Taylor approximation of interest, i.e., Eq. (E.3.17), in terms of $\tilde{p} \in \mathbb{N}_0$ and a subsequent sum over all subvectors $\vec{l}$ such that $\vec{l}_+ = \tilde{p}$.

Model selection consistency under the alternative To show that the Bayes factor increases under the alternative, irrespectively of γ_j being larger or smaller than 1, we study the exponential term of Eq. (6.2.2)

$$v(n) = \big(\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle\big)^{\frac{\boldsymbol{c}_+}{2} n} \prod_{j=1}^{K-1} \gamma_j^{-\frac{c_j}{2} n} \tag{E.3.30}$$

The claim is that v monotonically increases in n. Suppose that this is not true, then the ratio of subsequent terms

$$v(n+1)/v(n) = \big(\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle\big)^{\frac{\boldsymbol{c}_+}{2}} \prod_{j=1}^{K-1} \gamma_j^{-\frac{c_j}{2}} \tag{E.3.31}$$

would be less or equal to one. The gradient of $v(n+1)/v(n)$ with respect to γ is of the form

$$\tfrac{c_k}{2} \big(\tfrac{\boldsymbol{c}_+}{\langle \boldsymbol{c}, \boldsymbol{\gamma} \rangle} - \tfrac{1}{\gamma_k} \big) v(n+1)/v(n) \tag{E.3.32}$$

and this reveals a (global) minimum at $\boldsymbol{\gamma} = \mathbf{1}$ at which $v(n+1)/v(n) = 1$. Hence, any $\boldsymbol{\gamma} \neq \mathbf{1}$ leads to an exponentially increasing Bayes factor $\mathrm{BF}_{10}(\boldsymbol{s}^2, n)$. □

The proof of the previous theorem relies on a particular Lauricella function G_D to be of order 1 as n increases as shown in the following lemma.

Lemma 1 (Limit of a particular Lauricella function)**.** *For all $v_j, b_j > 0$, $j \in [m]$ and $|x_j| < 1$, we have that*

$$\lim_{n \to \infty} F_D(a\,;\, n\vec{v} + \vec{b}\,;\, v_+ n + b_+\,;\, \vec{x}) = \Big(1 - \sum_{i=1}^{m} \tfrac{v_i}{v_+} x_i\Big)^{-a}, \tag{E.3.33}$$

as $n \to \infty$. ◇

Proof. The proof follows from the asymptotic behavior of the gamma function combined with repeated use of the (negative) binomial series.

Firstly, note that the n dependence occurs in the lower and the upper terms of the second type, which cancels out as n grows large. To show this consider the definition of the Pochhammer raising factorial that combined with the Laplace approximation Eq. (E.3.19) for constants $v, b > 0$ leads to

$$(vn + b)_k = \frac{\Gamma(vn + b + k)}{\Gamma(vn + b)} = (vn)^k \big[1 + k(k + 2b - 1)(vn)^{-1} + \mathcal{O}((vn)^{-2})\big] \tag{E.3.34}$$

as $n \to \infty$.

Secondly, to describe the large n behavior of the particular type D Lauricella hypergeometric series $F_D := F_D(a\,;\, n\vec{v} + \vec{b}\,;\, v_+ n + b_+\,;\, \vec{x})$ we use the notation $i[k:m] = (i_j, \ldots, i_m) \in \mathbb{N}^{m-(k-1)}$ to denote the vector of indexes from k to m. Based on this notation and by Eq. (E.3.34), we have for n large that

$$\begin{aligned} F_D &= \sum_{i[1:m]} \frac{(a)_{i[1:m]_+} (v_1 n + b_1)_{i_1} \cdots (v_m n + b_m)_{i_m}}{(v_+ n + b_+)_{i[1:m]_+}} \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!} \qquad \text{(E.3.35)} \\ &\approx \sum_{i[1:m]} \frac{(a)_{i[1:m]_+} v_1^{i_1} \cdots v_m^{i_m}}{v_+^{i[1:m]_+}} \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!} = \sum_{i=\vec{0}}^{\infty} (a)_{i[1:m]_+} \frac{(\frac{v_1}{v_+} x_1)^{i_1}}{i_1!} \cdots \frac{(\frac{v_m}{v_+} x_m)^{i_m}}{i_m!}. \end{aligned}$$

The last equality defines the limit of F_D with respect to n. It also captures the essence of the repeated use of the binomial series, namely, the redistribution of the scaling factor $v_+^{-i[1:m]_+}$ over the variables x.

Thirdly, with the notation $i[2:m]$ it is simple to isolate the summation with respect to i_1 only, which combined with the binomial series yields

$$\lim F_D = \Big(\sum_{i_1=0}^{\infty} (a)_{i[1:m]_+} \frac{(\frac{v_1}{v_+} x_1)^{i_1}}{i_1!} \Big) \sum_{i[2:m]} \frac{(\frac{v_2}{v_+} x_2)^{i_2}}{i_2!} \cdots \frac{(\frac{v_m}{v_+} x_m)^{i_m}}{i_m!} \tag{E.3.36}$$

$$= \Big(\tfrac{v_+ - v_1 x_1}{v_+} \Big)^{-a} \sum_{i[2:m]} (a)_{i[2:m]_+} \Big(\tfrac{v_+ - v_1 x_1}{v_+} \Big)^{-i[2:m]_+} \frac{(\frac{v_2}{v_+} x_2)^{i_2}}{i_2!} \cdots \frac{(\frac{v_m}{v_+} x_m)^{i_m}}{i_m!}.$$

Note that, as before, the scaling factor $\Big(\tfrac{v_+ - v_1 x_1}{v_+} \Big)^{-i[2:m]_+}$ can be redistributed over the variables resulting in $(\frac{v_k}{v_+ - v_1 x_1} x_k)^{i_k} / i_k!$ for $k = 2, \ldots, m$. The summation with respect to i_2 is again a binomial series and yields

$$\lim F_D = \Big(\tfrac{v_+ - v_1 x_1}{v_+} \Big)^{-a} \Big(\tfrac{v_+ - v_1 x_1 - v_2 x_2}{v_+ - v_1 x_1} \Big)^{-a} \tag{E.3.37}$$

$$\times \sum_{i[3:m]} (a)_{i[3:m]_+} \Big(\tfrac{v_+ - v_1 x_1 - v_2 x_2}{v_+ - v_1 x_1} \Big)^{-i[3:m]_+} \frac{(\frac{v_3}{v_+ - v_1 x_1} x_3)^{i_3}}{i_3!} \cdots \frac{(\frac{v_m}{v_+ - v_1 x_1} x_m)^{i_m}}{i_m!}.$$

Observe that the numerator and denominator of the first and second $-a$ exponentiated terms in the previous display are equal and thus cancel. Repeating this procedure to m and telescoping through the $-a$ exponentiated terms yields the results. □

E.3.5 Limit and across-sample consistency

Proof of across-sample consistency, Theorem 6.2.5. To simplify notation we write $n := n_K$ and $\vec{\mathsf{ss}} = \overrightarrow{\nu s^2}$, where $\mathsf{ss}_j = \nu_j s_j^2$ is the sum of squares of the jth sample. Since S_K^2 is $\sqrt{n}$-consistent we can find an N such that for all $n > N$ the following statement holds with chance at least $1 - \epsilon$

$$\mathrm{BF}_{10}^{[K]}(\vec{s^2}, S_K^2, n) = \mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n) + \tfrac{h_n}{\sqrt{n}} \tilde{T}_1(n) + o_P(n^{-\frac{1}{2}}) \tilde{T}_2(n), \tag{E.3.38}$$

where h_n is a bounded sequence of random variables due to $S_K^2 - \sigma_0^2 = \mathcal{O}_P(n^{-\frac{1}{2}})$ and where

$$\tilde{T}_1(n) = \Big(\tfrac{\partial}{\partial x} \mathrm{BF}_{10}^{[K]}(\vec{s^2}, x, n) \Big) \Big|_{x=\sigma_0^2}, \tag{E.3.39}$$

$$\tilde{T}_2(n) = \Big(\tfrac{\partial^2}{\partial x^2} \mathrm{BF}_{10}^{[K]}(\vec{s^2}, x, n) \Big) \Big|_{x=\sigma_0^2}. \tag{E.3.40}$$

To prove the theorem we have to show that $\lim_{n \to \infty} \mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n)$ exists, is equal to Eq. (6.2.9), and that both $\tilde{T}_1(n)$ and $\tilde{T}_2(n)$ are bounded in n. To this end, we want to first take the limit and then integrate. To see that this is permissible we first show that the integrand of $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n)$ as a sequence in n is uniformly bounded in $\vec{\vartheta}$.

Uniformly boundedness of the integrand To further simplify notation we introduce the vectors $\vec{a}, \vec{c} \in \mathbb{R}^{K-1}$ with $a_j = \frac{\nu_j}{2}$ for $j \in [K-1]$ and $b = \frac{n}{2}$. By definition of $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n)$, the innocuous replacement $n = \nu_K$ we have that

$$\mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n) = (1 + \tfrac{\mathsf{s}\vec{\mathsf{s}}_+}{n\sigma_0^2})^{a_+ + b} \int \tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta}) \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta}, \tag{E.3.41}$$

where $\pi_1(\vec{\vartheta})$ is the Dirichlet prior with parameters $\boldsymbol{u}$ and where

$$\tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta}) = \Big(\prod_{j=1}^{K-1} \vartheta_j^{a_j} \Big) (1 - \vec{\vartheta}_+)^b (1 - \sum_{j=1}^{K-1} [1 - \tfrac{\mathsf{ss}_j}{n\sigma_0^2}] \vartheta_j)^{-(a_+ + b)}, \tag{E.3.42}$$

is the marginalized likelihood with σ_0^2 in place of s_K^2, thus, $\tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta}_0) = (1 + \frac{\mathsf{s}\vec{\mathsf{s}}_+}{n\sigma_0^2})^{-(a_+ + b)}$. By definition of the exponential function as a series, the first term in Eq. (E.3.41) remains bounded, that is,

$$\lim_{n \to \infty} (1 + \tfrac{\mathsf{s}\vec{\mathsf{s}}_+}{n\sigma_0^2})^{\frac{\vec{\nu}_+ + n}{2}} = e^{\frac{\mathsf{s}\vec{\mathsf{s}}_+}{2\sigma_0^2}} \Big(1 - \tfrac{\mathsf{s}\vec{\mathsf{s}}_+}{4n\sigma_0^2} (\tfrac{\mathsf{s}\vec{\mathsf{s}}_+}{\sigma_0^2} - 2\vec{\nu}_+) + \mathcal{O}(n^{-2}) \Big). \tag{E.3.43}$$

The prior does not play a role in the asymptotics for $n \to \infty$, as we will show that

$$\int \tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta}) \pi_1(\vec{\vartheta}) \mathrm{d}\vec{\vartheta} \leq C(\boldsymbol{u}) \int \tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta}) \mathrm{d}\vec{\vartheta}. \tag{E.3.44}$$

for a certain constant $C(\boldsymbol{u})$ independent of n.

Case (i) The case with $\boldsymbol{u}$ all at least 1, we can take $C(\boldsymbol{u})$ to be the maximum of the prior $\mathrm{Dir}(\vec{\vartheta}; \boldsymbol{u})$ on $\vec{\vartheta}$ in the $K-1$ simplex. The maximum of the marginalized likelihood $\tilde{h}(\vec{s^2}, \sigma_0^2, n \,|\, \vec{\vartheta})$ at each n can be found by setting the partial derivatives to zero. At each fixed n Lemma 2 can be used to find the maximum $\hat{\vartheta}$ as a function of $\vec{a}, b, \vec{c}$. By definition of $\vec{a}, b, \vec{c}$ and by denoting the observed precisions $\vec{t} \in \mathbb{R}^{K-1}$ by $t_j := (s_j^2)^{-1}$, it then follows that $\hat{\vartheta}_k = \frac{t_k}{\sigma_0^{-2} + \vec{t}_+}$, which is free of n. A plugin and a direct calculation show that the maximum value of the marginalized likelihood at each n is

$$f_{\max,n} := \Big(\prod_{k=1}^{K-1} t_k^{\frac{\nu_k}{2}} \Big) (\sigma_0^2)^{\frac{\vec{\nu}_+}{2}} e^{-\frac{\vec{\nu}_+}{2}} [1 - \tfrac{\vec{\nu}_+^2}{4n} + \mathcal{O}(n^{-2})]. \tag{E.3.45}$$

Hence, as a sequence in n the integrand is uniformly bounded by a constant.

Case (ii) For any $u_j < 1$, $j \in [K-1]$ the prior diverges at $\vartheta_j = 0$ and $C(\boldsymbol{u})$ cannot be taken to be the maximum value of the prior on the $K-1$ simplex. Instead, $C(\boldsymbol{u})$ can be the maximum of $\pi_1(\vec{\vartheta})$ for $\vec{\vartheta}$ in a subset R containing $\hat{\vartheta}$. Since the true variances are assumed to be non-zero, finite and the data continuous, we can take R with high probability to be a compact subset that intersects with $\bigoplus_{j=1}^{K-1}[\epsilon_j, 1-\epsilon_j] \subset [0,1]^{K-1}$ for ϵ_j depending on u_j. On R the proof of Case (i) can be repeated to show that that the integrand is bounded. For any $u_j < 1$, $j \in [K-1]$ the integrand over $\vartheta_j \in [0, \epsilon_j)$ behaves as $\vartheta_j^{\frac{\nu_j}{2}+u_j-1} + \mathcal{O}(|\vartheta_j|)$. On this domain the integrand remains integrable whenever $u_j > -\frac{\nu_j}{2}$, which is true by assumption. The same arguments extend to the case with $u_K < 1$.

Identifying the $K-1$-sample Bayes factor Uniform boundedness allows us to interchange the limit and integral and conclude that the limiting integral exist, and implies that $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, \sigma_0^2, n)$ converges to

$$\frac{\int \Big(\prod \vartheta_j^{\frac{\nu_j}{2}+u_j-1}\Big)(1-\vec{\vartheta}_+)^{u_K - \frac{\vec{\nu}_+}{2}-1} \exp\big(-\sum \frac{\mathrm{ss}_j}{2\sigma_0^2}(\frac{\vartheta_j}{1-\vec{\vartheta}_+})\big)\mathrm{d}\vec{\vartheta}}{\mathcal{B}(\boldsymbol{u})\exp(-\frac{\vec{\mathrm{ss}}_+}{2\sigma_0^2})}. \tag{E.3.46}$$

From the change of variables $\vartheta_j = \frac{\xi_j}{1+\xi_+}$, thus, $\mathrm{d}\vec{\vartheta} = (1+\xi_+)^{-K}\mathrm{d}\vec{\xi}$, and by definition of the integral representation of the multivariable Tricomi function $\mathcal{U}$, see for instance (Ng et al., 2011; Phillips, 1988), we have that the resulting $K-1$ sample Bayes factor is given by

$$\begin{aligned}\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2}) &= \frac{\int \Big(\prod_{j=1}^{K-1}\tau_j^{\frac{\nu_j}{2}}\Big)\exp(-\frac{1}{2}\sum_{j=1}^{K-1}\nu_j s_j^2\tau_j)\pi_{\sigma_0^2}(\vec{\tau}\,|\,\mathcal{M}_1^{[K-1]})\mathrm{d}\vec{\tau}}{(\sigma_0^2)^{-\frac{\vec{\nu}_+}{2}}\exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})}, \\ &= \frac{\Big(\prod_{j=1}^{K-1}\Gamma(\frac{\nu_j}{2}+u_j)\Big)\mathcal{U}\Big(\frac{\vec{\nu}}{2}+\vec{u}\,;\,\frac{\vec{\nu}_+}{2}-u_K+1\,;\,\frac{\overrightarrow{\nu s^2}}{2\sigma_0^2}\Big)}{\mathcal{B}(\vec{u},w)\exp(-\frac{(\overrightarrow{\nu s^2})_+}{2\sigma_0^2})},\end{aligned} \tag{E.3.47}$$

where $\overrightarrow{\nu s^2} = (\nu_1 s_1^2, \ldots, \nu_{K-1}s_{K-1}^2)$ denotes the vector of sums of squares, $(\overrightarrow{\nu s^2})_+ = \sum_{j=1}^{K-1}\nu_j s_j^2$, and $\vec{\nu}_+ := \sum_{j=1}^{K-1}\nu_j$, as before. This Bayes factor is based on uniform priors on the nuisance parameters $\vec{\mu} \in \mathbb{R}^{K-1}$, and an inverse Dirichlet distribution on the precisions $\vec{\tau} = (\tau_1, \ldots, \tau_{K-1}) \in \mathbb{R}^{K-1}$ scaled by $1/\sigma_0^{-2}$, that is,

$$\pi_{\sigma_0^2}(\vec{\tau}\,|\,\mathcal{M}_1^{[K-1]}) = \frac{(\sigma_0^2)^{K-1}\prod_{j=1}^{K-1}(\sigma_0^2\tau_j)^{u_j-1}}{\mathcal{B}(\vec{u},w)(1+\sigma_0^2\vec{\tau}_+)^{\vec{u}_+ + w}}, \tag{E.3.48}$$

where we wrote $w = u_K$ so the statement only involves vectors of length $K-1$.

Recall that $\texttt{ss}_j = \nu_j s_j^2$ summarizes the observations of the jth sample. Observe also that the numerator of this limiting Bayes factor resembles the marginalized likelihood, i.e., Eq. (E.2.3), of the $K-1$ samples with their respective precisions $\vec{\tau} = (\tau_1, \ldots, \tau_{K-1})$ all fixed at $1/\sigma_0^2$. Hence, up to the factor $(\sigma_0^2)^{-\frac{\vec{\nu}_+}{2}}$ the denominator defines the marginal likelihood of the lower-dimensional null hypothesis $\mathcal{H}_0^{K-1} : \tau_j = \sigma_0^{-2}$ for $j \in [K-1]$ with $\mu_j \propto 1$. The missing factor is retrieved from the numerator by the change of variable $\tau_j = \frac{\vartheta_j}{\sigma_0^2(1-\vec{\vartheta}_+)}$ and yields the assertion above Eq. (6.2.8).

The lower dimensional Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ is in general hard to compute, because the Tricomi function $\mathcal{U}(\vec{b}\,;\, c\,;\, \vec{x})$ defines a $K-1$-dimensional integral. Phillips (1988) showed that if $c < 1$, the following simplification holds

$$\mathcal{U}(\vec{b}\,;\, c\,;\, \vec{x}) = \int_0^\infty e^{-t} t^{\vec{b}_+ - c} \prod_{j=1}^{K-1} (t + x_j)^{-b_j} \mathrm{d}t. \tag{E.3.49}$$

For $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ this simplification holds whenever $\vec{\nu}_+ < 2u_K$, which will be of little practical use when, for instance, $u_K = 1/2$. Theorem 6.2.5 now shows that for the case with $\vec{\nu}_+ \geq 2u_K$ the lower dimensional Bayes factor $\mathrm{BF}_{10\,;\,\sigma_0^2}^{[K-1]}(\vec{s^2})$ can be well approximated by a one-dimensional integral, because the type D Lauricella function in $\mathrm{BF}_{10}(\boldsymbol{s^2})$ has a simplified one-dimensional integral representation due to $\boldsymbol{u}_+ > 0$.

Residual terms To show that the convergence is at rate $1/\sqrt{n}$, we show that both $\tilde{T}_1(n)$ and $\tilde{T}_2(n)$ in Eq. (E.3.38) are of order 1. The analysis is analogous to showing the existence of $\mathrm{BF}_{10}^{[K-1]}$.

For $\tilde{T}_1(n)$ we study the derivative of the Bayes factor $\mathrm{BF}_{10}^{[K]}(\vec{s^2}, x, n)$ with respect to x. For this we swap the order of integration and differentiation and consider

$$g := \tfrac{\partial}{\partial x} \tfrac{h(\vec{s^2}, x, n \,|\, \vec{\vartheta})}{h(\vec{s^2}, x, n \,|\, \vec{\vartheta}_0)} \Big|_{x=\sigma_0^2} = g_1 + g_2 \tag{E.3.50}$$

where

$$g_1 := -\tfrac{\vec{\mathsf{ss}}_+}{n\sigma_0^4}(a_+ + b)(1 + \tfrac{\vec{\mathsf{ss}}_+}{n\sigma_0^2})^{a_+ + b - 1} \tag{E.3.51}$$

$$\times \Big(\prod_{j=1}^{K-1} \vartheta_j^{a_j} \Big)(1 - \vec{\vartheta}_+)^b (1 - \sum_{j=1}^{K-1} [1 - \tfrac{\mathsf{ss}_j}{n\sigma_0^2}]\vartheta_j)^{-(a_+ + b)}, \tag{E.3.52}$$

$$g_2 := \tfrac{1}{n\sigma_0^4}(a_+ + b)(1 + \tfrac{\vec{\mathsf{ss}}_+}{n\sigma_0^2})^{a_+ + b} \tag{E.3.53}$$

$$\sum_{k=1}^{K-1} \mathsf{ss}_k \vartheta_k \Bigg[\Big(\prod_{j=1}^{K-1} \vartheta_j^{a_j} \Big)(1 - \vec{\vartheta}_+)^b (1 - \sum_{j=1}^{K-1} [1 - \tfrac{\mathsf{ss}_j}{n\sigma_0^2}]\vartheta_j)^{-a_+ - 1 - b} \Bigg]. \tag{E.3.54}$$

Note that by definition of $\vec{a}, b, \vec{\mathsf{ss}}$ the terms Eq. (E.3.51) and Eq. (E.3.53) converge to $-\frac{\vec{\mathsf{ss}}_+}{2\sigma_0^4} e^{\frac{\vec{\mathsf{ss}}_+}{2\sigma_0^2}}$ and $\frac{1}{2\sigma_0^4} e^{\frac{\vec{\mathsf{ss}}_+}{2\sigma_0^2}}$, respectively. The proof that Eq. (E.3.52) is uniformly bounded in n is exactly as before. The same proof holds for each member in the sum of Eq. (E.3.54) by relabelling the power corresponding to ϑ_k to $a_k + 1$. Hence, limit and integral can be interchanged and we conclude that the limiting integral exists. A computation as before shows that

$$\breve{T}_1 := \lim_{n \to \infty} \Big(\tfrac{\partial}{\partial x} \mathrm{BF}_{10}(\vec{s^2}, x, n) \Big)\Big|_{x = \sigma_0^2} = \frac{\prod_{j=1}^{K-1} \Gamma(\frac{\nu_j}{2} + u_j)}{2\mathcal{B}(\boldsymbol{u})\sigma_0^4 \exp(-\frac{\vec{\mathsf{ss}}_+}{2\sigma_0^2})} G_2, \tag{E.3.55}$$

where

$$G_2 := \sum_{k=1}^{K-1} (\tfrac{\nu_k}{2} + u_k)\mathcal{U}(\tfrac{\vec{\nu}}{2} + \vec{u} + \vec{e}_k \,;\, \tfrac{\vec{\nu}_+ + 1}{2} - w + 1 \,;\, \tfrac{\vec{\mathsf{ss}}}{2\sigma_0^2}) \tag{E.3.56}$$

$$- \vec{\mathsf{ss}}_+ \mathcal{U}(\tfrac{\vec{\nu}}{2} + \vec{u} \,;\, \tfrac{\vec{\nu}_+}{2} - w + 1 \,;\, \tfrac{\vec{\mathsf{ss}}}{2\sigma_0^2}), \tag{E.3.57}$$

where $\vec{e}_k \in \mathbb{R}^{K-1}$ denotes the kth basis vector that is one at the kth entry and zero elsewhere. The analysis of the third order term is a repeat of that of $\breve{T}_1$ and implies that the last term in Eq. (E.3.38) is indeed $o_P(n^{-\frac{1}{2}})$ and the result follows. □

If Y_{Ki} has four moments, then S_K^2 is asymptotically normal. In particular, for normal data this explicitly means $\sqrt{n}(S_K^2 - \tau^{-1}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 2\tau^{-2})$ and implies the following result.

Proof of asymptotic normality across-samples. A rewrite of Eq. (E.3.41) shows that n_K

$$\sqrt{n_K}\Big(\mathrm{BF}_{10}^{[K]}(y^{[K]}) - \mathrm{BF}_{10\,;\,\sigma_K^2}^{[K-1]}(y^{[K-1]})\Big) = \sqrt{n_K}\Big(S_K^2 - \tau^{-1}\Big)\tilde{T}_2(n_K) \tag{E.3.58}$$

$$+ o_P(1)\tilde{T}_3(n_K). \tag{E.3.59}$$

A series expansion of $\tilde{T}_2(n_K)$ in n_K shows that $\tilde{T}_2(n_K) = T_2 + \frac{1}{n}\breve{T}_2 + \mathcal{O}(\frac{1}{n_K^2})$ and the result follows. The term $\breve{T}_2$ can be derived explicitly as was done in the proof of the previous theorem, but does not matter for the assertion, but its presence reveals a finite sample $\mathcal{O}(n_K^{-1/2})$ bias that vanishes as $n_K \to \infty$. □

The proof of across sample consistency relies on the following lemma.

Lemma 2 (Maximum of the marginalized likelihood). *If $\vec{a}, \vec{c}, \vec{\vartheta} \in \mathbb{R}^{K-1}$ and $b \in \mathbb{R}$ all positive and $\vec{\vartheta}_+ < 1$, then*

$$f(\vec{a}, b, \vec{c} \,|\, \vec{\vartheta}) = \Big(\prod_{j=1}^{K-1} \vartheta_j^{a_j} \Big) (1 - \vec{\vartheta}_+)^b (1 - \sum_{j=1}^{K-1} [1 - c_j] \vartheta_j)^{-(a_+ + b)}, \tag{E.3.60}$$

attains its maximum at

$$\hat{\vartheta}_k = \frac{a_k \prod_{j \neq k}^{K-1} c_j}{b \prod_{j=1}^{K-1} c_j + \sum_{i=1}^{K-1} a_i \prod_{j \neq i}^{K-1} c_j}, \tag{E.3.61}$$

where $\prod_{j \neq k}^{K-1} c_j$ denotes the product of the elements of $\vec{c}$ with the kth element taken out. ◇

Proof. Recall that the maximum is invariant under smooth transformations, which allows us to study the problem in the parametrisation $\vec{\xi} = (\xi_1, \ldots, \xi_{K-1})$, where $\vartheta_j = \frac{\xi_j}{1 + \vec{\xi}_+}$. The target function becomes

$$f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) = \Big(\prod_{j=1}^{K-1} \xi_j^{a_j} \Big) (1 + \sum_{j=1}^{K-1} c_j \xi_j)^{-(a_+ + b)}, \tag{E.3.62}$$

and a direct computation shows that its gradient consists of elements

$$\tfrac{\partial}{\partial \xi_k} f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) = f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) \Big[\tfrac{a_k}{\xi_k} - \tfrac{(a_+ + b) c_k}{1 + \sum_{j=1}^{K} c_j \xi_j} \Big]. \tag{E.3.63}$$

It is now easy to verify that for $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_{K-1})$ with $\hat{\xi}_k = \frac{a_k}{b c_k}$ the vector of partial derivatives is zero. Straightforward calculations show that for $k \neq l \in [K-1]$ that

$$\tfrac{\partial^2}{\partial \xi_k \partial \xi_l} f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) = f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) \Big[\tfrac{a_k}{\xi_k} - \tfrac{(a_+ + b) c_k}{1 + \sum_{j=1}^{K} c_j \xi_j} \Big]_{\vec{\xi} = \hat{\xi}} = \tfrac{b^2 c_k c_l}{a_+ + b} \tag{E.3.64}$$

and for $k \in [K-1]$

$$\tfrac{\partial^2}{\partial \xi_k^2} f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) = f(\vec{a}, b, \vec{c} \,|\, \vec{\xi}) \Big[\tfrac{a_k}{\xi_k} - \tfrac{(a_+ + b) c_k}{1 + \sum_{j=1}^{K} c_j \xi_j} \Big]_{\vec{\xi} = \hat{\xi}} = - \tfrac{(b c_k)^2 (a[-k]_+ + b)}{a_k (a_+ + b)}, \tag{E.3.65}$$

from which we conclude that $\hat{\xi}$ is a maximum. The transformation $\hat{\vartheta} = \frac{\hat{\xi}_k}{1 + \hat{\xi}_+}$ yields the results. □

E.4 Analysis Code

Here, we provide the code for all examples given in the main text.

```
devtools::install_github('fdabl/bfvartest', build_vignettes = TRUE)
library('bfvartest')

# 5.1 Sex Differences in Personality
twosd_test(n1 = 969, n2 = 716, sd1 = sqrt(15.6),
           sd2 = sqrt(19.9), u = 0.50)

# 5.2 Testing Against a Single Value
x <- c(6.2, 5.8, 5.7, 6.3, 5.9, 5.8, 6.0)
n <- length(x)
sd_x <- sd(x) # use rounded 0.22 in the paper

## (i) BF_{+0}
onesd_test(
    n = n, s = sd_x, popsd = sqrt(0.10),
    u = 0.50, alternative_interval = c(1, Inf), log = FALSE
)

## (ii) BF_{10}
onesd_test(
    n = n, s = sd_x, popsd = sqrt(0.10),
    u = 0.50, alternative_interval = c(0, Inf), log = FALSE
)

## (iii) BF_{+0} informed
onesd_test(
    n = n, s = sd_x, popsd = sqrt(0.10),
    u = 2.16, alternative_interval = c(1, Inf), log = FALSE
)

# 5.3 Comparing Measurement Precision
n <- 990
sdigit <- 0.98
slaser <- 0.89

## (i) BF_{+0}
twosd_test(
    n1 = n, n2 = n, sd1 = slaser, sd2 = sdigit,
    u = 0.50, alternative_interval = c(1, Inf), log = FALSE
)

## (ii) BF'_{0+} non-overlapping interval
1 / twosd_test(
    n1 = n, n2 = n, sd1 = slaser, sd2 = sdigit, u = 0.50,
    log = FALSE, null_interval = c(0.90, 1.10),
    alternative_interval = c(1.10, Inf)
)

# 5.4 The "Standardization" Hypothesis in Archeology
ns <- c(117, 171, 55)
sds <- c(12.74, 8.13, 5.83)
hyp <- c('1=2=3', '1,2,3', '1>2>3')
res <- ksd_test(hyp = hyp, ns = ns, sds = sds, u = 0.50,
                iter = 6000)
res$BF

# 5.5 Increased Variability in Mathematical Ability
ns <- c(3280, 6007, 7549, 9160, 9395, 6410)
sds <- c(5.99, 5.39, 4.97, 4.62, 3.69, 3.08)
hyp <- c('1=2=3=4=5=6', '1,2,3,4,5,6', '1>2>3>4>5>6')
res <- ksd_test(hyp = hyp, ns = ns, sds = sds, u = 0.50,
                iter = 6000)
res$BF
```

F
Appendix of Chapter 7

F.1 Example Code

The code below illustrates the proportion example in subsection 7.4.1. To install the package enter the Pkg REPL by typing] and `add EqualitySampler`. Alternatively, the package can be installed by importing the Pkg package: `import Pkg; Pkg.add(EqualitySampler)`.

```
using EqualitySampler, EqualitySampler.Simulations
import DataFrames      as DF,
       LinearAlgebra   as LA,
       NamedArrays     as NA,
       CSV,
       AbstractMCMC

# assumes the working directory is the root of the GitHub repository
journal_data = DF.DataFrame(CSV.File(joinpath("simulations", "demos",
    "data", "journal_data.csv")))

# K
n_journals = size(journal_data, 1)
# no of observed errors
errors = round.(Int, journal_data.n .* journal_data.errors)
# no of possible errors
observations = journal_data.n

# run 4 chains in parallel with 15_000 iterations per chain of which
# the first 5_000 are discarded
mcmc_settings = MCMCSettings(;iterations = 15_000, burnin = 5_000,
    chains = 4, parallel = AbstractMCMC.MCMCThreads)

# nothing indicates no equality sampling is done and samples are
# drawn from the full model
chn_full = proportion_test(errors, observations, nothing;
    mcmc_settings = mcmc_settings)
```

```
# use a BetaBinomial(1, k) over the partitions
partition_prior = BetaBinomialPartitionDistribution(n_journals,
    1, n_journals)
chn_eqs  = proportion_test(errors, observations, partition_prior;
mcmc_settings = mcmc_settings)
# chn_full and chn_eqs contain posterior samples

# the posterior probability that two journals are equal
eqs_mat = compute_post_prob_eq(chn_eqs)
NA.NamedArray(
    LA.UnitLowerTriangular(round.(eqs_mat; digits = 3)),
    (journal_data.journal, journal_data.journal)
)
# 8×8 Named LinearAlgebra.UnitLowerTriangular{Float64, Matrix{Float64}}
#  A \ B │  "JAP"    "PS"  "JCCP"  "PLOS"    "FP"    "DP"  "JEPG"  "JPSP"
# ───────┼───────────────────────────────────────────────────────────────
# "JAP"  │    1.0     0.0     0.0     0.0     0.0     0.0     0.0     0.0
# "PS"   │  0.134     1.0     0.0     0.0     0.0     0.0     0.0     0.0
# "JCCP" │    0.0     0.0     1.0     0.0     0.0     0.0     0.0     0.0
# "PLOS" │    0.0     0.0   0.909     1.0     0.0     0.0     0.0     0.0
# "FP"   │    0.0     0.0   0.861    0.87     1.0     0.0     0.0     0.0
# "DP"   │    0.0     0.0   0.864   0.886   0.881     1.0     0.0     0.0
# "JEPG" │    0.0     0.0   0.059   0.063    0.09    0.08     1.0     0.0
# "JPSP" │    0.0     0.0     0.0     0.0   0.005     0.0   0.852     1.0
# The table above is approximately equal to the right panel of Figure 7.6
```

F.2 Beta-binomial Prior with Decreasing Prior Model Odds

Proposition 3. *The prior density of the beta-binomial distribution over partitions is decreasing for $\alpha = 1$ and $\beta \geq \binom{K}{2}$, and strictly decreasing for $\alpha = 1$ and $\beta > \binom{K}{2}$.*

Proof. The prior density of the Beta-binomial over partitions is given by:

$$\pi\left(\rho \mid K, \alpha, \beta\right) = \binom{K-1}{|\rho|-1} \frac{\mathrm{B}\left(|\rho| - 1 + \alpha,\ K - |\rho| + \beta\right)}{\mathrm{B}\left(\alpha,\ \beta\right) \left\{ {K \atop |\rho|} \right\}} \; .$$

To examine the ratio of two consecutive model sizes we evaluate the ratio of the prior for partitions ρ and q with $|q| = |\rho| + 1$:

$$\frac{\pi\left(\rho \mid K,\, \alpha,\, \beta\right)}{\pi\left(q \mid K,\, \alpha,\, \beta\right)} = \frac{\binom{K-1}{|\rho|-1}}{\binom{K-1}{|\rho|}} \frac{\mathrm{B}\left(|\rho| - 1 + \alpha,\ K - |\rho| + \beta\right)}{\mathrm{B}\left(|\rho| + \alpha,\ K - |\rho| - 1 + \beta\right)} \frac{\left\{ {K \atop |\rho|+1} \right\}}{\left\{ {K \atop |\rho|} \right\}}, \tag{F.2.1}$$

$$= \frac{|\rho|}{K - |\rho|} \frac{\beta + K - |\rho| - 1}{\alpha + |\rho| - 1} \frac{\left\{ {K \atop |\rho|+1} \right\}}{\left\{ {K \atop |\rho|} \right\}}. \tag{F.2.2}$$

Using the recurrence relation of the Stirling numbers $\left\{{n+1 \atop k}\right\} = k\left\{{n \atop k}\right\} + \left\{{n \atop k-1}\right\}$, the ratio $\left\{{K \atop |\rho|+1}\right\}/\left\{{K \atop |\rho|}\right\}$ is equivalent to $\left\{{K+1 \atop |\rho|+1}\right\}/\left\{{K \atop |\rho|+1}\right\} - (|\rho| + 1)$. This ratio of Stirling numbers was studied by Berg (1975) and their property 2 provides this inequality

$$\frac{\left\{{K+1 \atop |\rho|+1}\right\}}{\left\{{K \atop |\rho|+1}\right\}} - |\rho| - 1 \geq \frac{\left\{{K+1 \atop |\rho|}\right\}}{\left\{{K \atop |\rho|}\right\}} - |\rho|.$$

It follows that the ratio in Equation (F.2.1) is maximal at $|\rho| = K - 1$ and has value $\binom{K}{2}$. Next, we fix $\alpha = 1$ and solve ${\pi(K|K-1,1,\beta)}/{\pi(K|K,1,\beta)} = 1$ for β which yields $\beta = \binom{K}{2}$. Thus $\beta \geq \binom{K}{2}$ implies $\pi\,(j+1 \mid K,\,1,\,\beta) \geq \pi\,(j \mid K,\,1,\,\beta)$ (resp. $\beta > \binom{K}{2}$ implies $\pi\,(j+1 \mid K,\,1,\,\beta) > \pi\,(j \mid K,\,1,\,\beta)$). □

F.3 Simulation Results for $K = 9$

Here we present the extended simulation results for the $K = 9$ group case. Figure F.1 mirrors the results for the $K = 9$ case, namely that the pairwise Bayes factors, the method proposed by Westfall et al. (1997), and the uniform prior generally increase in performance as the number of inequalities increase, while the other priors generally decrease in performance. Averaging over the settings, we again find that the beta-binomial prior with $\beta = 1$, the uniform prior, and the symmetric DP prior exhibit the worst error control, with the method proposed by Westfall et al. (1997) performing best, closely followed by the beta-binomial prior with $\beta = \binom{K}{2}$ and the DP prior with $\alpha = 0.50$.

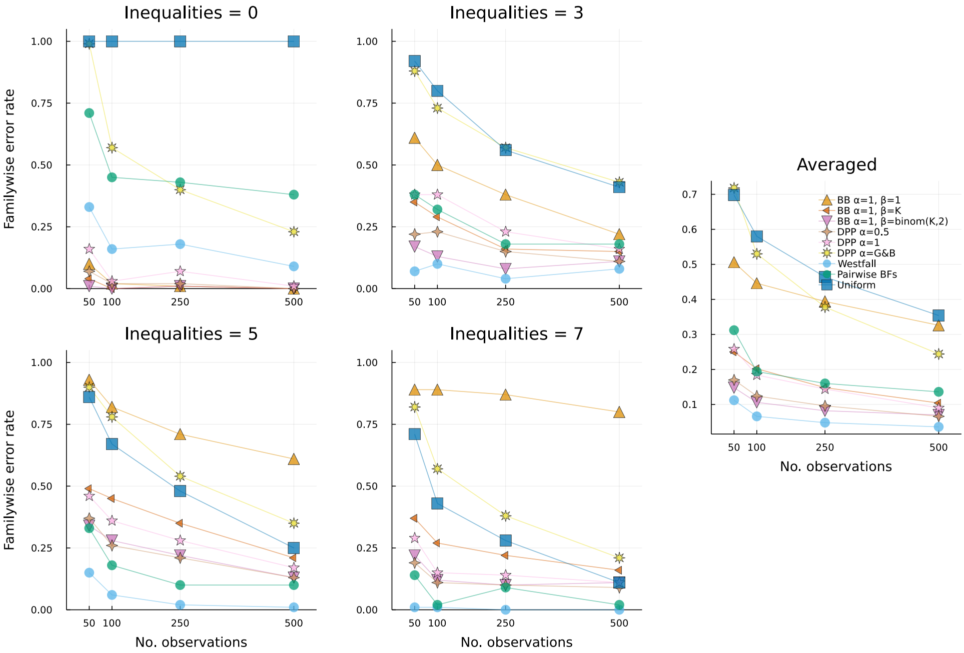


Figure F.1: Familywise error rate across priors and sample sizes under a model with 0 (top left), 3 (top right), 5 (bottom left), and 7 (bottom right) true inequalities for $K = 9$ groups. The rightmost panel shows the average familywise error rate across inequalities.

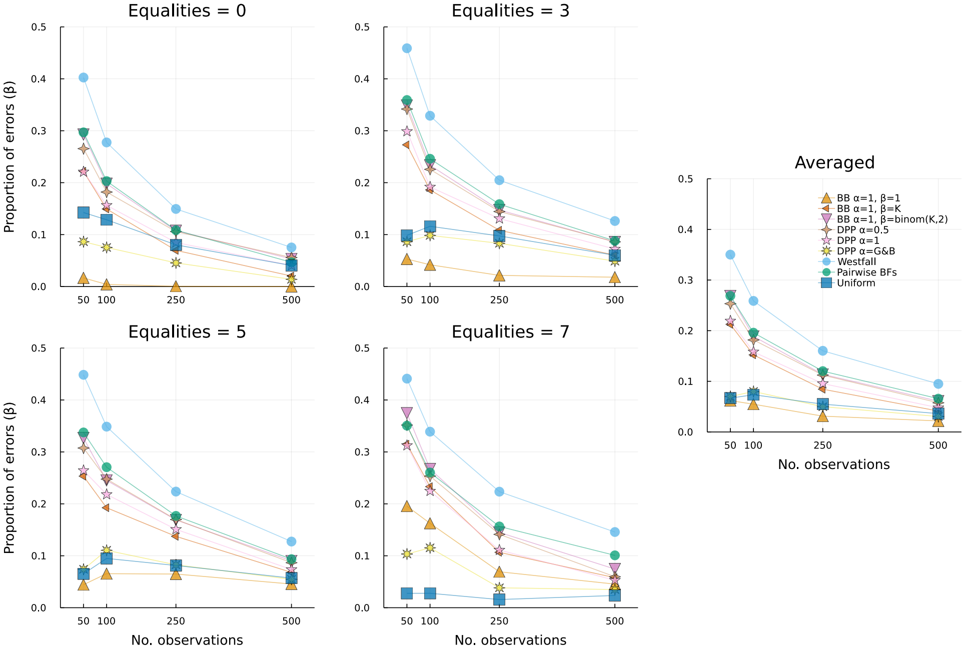


Figure F.2: Proportion of falsely claiming a difference between two groups when there is none across priors and sample sizes under a model with 0 (top left), 3 (top right), 5 (bottom left), and 7 (bottom right) true inequalities for $K = 9$ groups. The rightmost panel shows the average error rate across inequalities.

G
Appendix of Chapter 8

G.1 Complete Model Comparison Table

Table G.1: Bayesian multi-model inference for the World Happiness example: all 80 models. The leftmost column gives the model specification; the second column gives the prior model probabilities; the third the posterior model probabilities; the fourth the change from prior odds to posterior odds; the fifth the Bayes factor relative to the best model; and the last gives R^2.

Models	$P(\mathcal{M})$	$P(\mathcal{M} \mid \text{data})$	$\text{BF}_{\mathcal{M}}$	BF_{10}	R^2
W + Le + Ss + F + Le * Ss	0.013	0.759	248.244	1.000	0.821
W + Le + Ss + F + Ge + Le * Ss	0.013	0.097	8.531	7.783	0.822
W + Le + Ss + F + Poc + Le * Ss	0.013	0.093	8.101	8.157	0.822
Le + Ss + F + Le * Ss	0.013	0.027	2.233	27.591	0.805
W + Le + Ss + F + Ge + Poc + Le * Ss	0.013	0.012	0.924	65.617	0.823
Le + Ss + F + Ge + Le * Ss	0.013	0.005	0.413	145.922	0.807
Le + Ss + F + Poc + Le * Ss	0.013	0.004	0.329	182.965	0.807
W + Le + Ss + F	0.013	6.961×10^{-4}	0.055	1089.774	0.794
Le + Ss + F + Ge + Poc + Le * Ss	0.013	6.672×10^{-4}	0.053	1137.027	0.808
W + Le + Ss + F + Poc	0.013	3.179×10^{-4}	0.025	2386.195	0.799
W + Le + Ss + F + Ge	0.013	2.676×10^{-4}	0.021	2834.341	0.799
W + Ss + F	0.013	1.216×10^{-4}	0.010	6239.227	0.781
W + Le + Ss + Poc + Le * Ss	0.013	8.133×10^{-5}	0.006	9327.093	0.795
W + Le + Ss + F + Ge + Poc	0.013	6.763×10^{-5}	0.005	11 216.690	0.802
W + Le + Ss + Ge + Le * Ss	0.013	6.430×10^{-5}	0.005	11 796.826	0.794
W + Ss + F + Poc	0.013	5.121×10^{-5}	0.004	14 813.739	0.786
W + Le + Ss + Le * Ss	0.013	4.945×10^{-5}	0.004	15 340.968	0.786
W + Ss + F + Ge	0.013	4.745×10^{-5}	0.004	15 988.688	0.786
W + Le + Ss + Ge + Poc + Le * Ss	0.013	2.911×10^{-5}	0.002	26 057.578	0.799
Le + Ss + Ge + Le * Ss	0.013	1.404×10^{-5}	0.001	54 049.136	0.782
Le + Ss + Poc + Le * Ss	0.013	1.313×10^{-5}	0.001	57 757.710	0.782
W + Ss + F + Ge + Poc	0.013	1.102×10^{-5}	8.710×10^{-4}	68 808.309	0.789
Le + Ss + F	0.013	8.251×10^{-6}	6.518×10^{-4}	91 942.898	0.772
Le + Ss + F + Ge	0.013	8.136×10^{-6}	6.427×10^{-4}	93 244.135	0.780
Le + Ss + F + Poc	0.013	7.467×10^{-6}	5.899×10^{-4}	101 586.552	0.780
Le + Ss + Ge + Poc + Le * Ss	0.013	6.790×10^{-6}	5.364×10^{-4}	111 729.632	0.787
Le + Ss + Le * Ss	0.013	5.554×10^{-6}	4.388×10^{-4}	136 585.291	0.771
W + Le + F	0.013	2.891×10^{-6}	2.284×10^{-4}	262 420.104	0.769
Le + Ss + F + Ge + Poc	0.013	2.704×10^{-6}	2.136×10^{-4}	280 537.628	0.784
W + Le + F + Ge	0.013	9.872×10^{-7}	7.799×10^{-5}	768 432.339	0.773
W + Le + F + Poc	0.013	6.255×10^{-7}	4.941×10^{-5}	1.213×10^{6}	0.772
W + Le + Ss + Poc	0.013	4.229×10^{-7}	3.341×10^{-5}	1.794×10^{6}	0.770
W + Le + Ss + Ge + Poc	0.013	4.004×10^{-7}	3.163×10^{-5}	1.894×10^{6}	0.778
W + F	0.013	2.744×10^{-7}	2.168×10^{-5}	2.764×10^{6}	0.751
W + Le + Ss + Ge	0.013	1.846×10^{-7}	1.459×10^{-5}	4.109×10^{6}	0.768
W + Le + F + Ge + Poc	0.013	1.459×10^{-7}	1.152×10^{-5}	5.200×10^{6}	0.775
W + F + Ge	0.013	9.281×10^{-8}	7.332×10^{-6}	8.174×10^{6}	0.757

Models	$P(\mathcal{M})$	$P(\mathcal{M} \mid \text{data})$	$\text{BF}_{\mathcal{M}}$	BF_{01}	R^2
Le + Ss + Ge + Poc	0.013	6.433×10^{-8}	5.082×10^{-6}	1.179×10^{7}	0.764
W + F + Poc	0.013	5.171×10^{-8}	4.085×10^{-6}	1.467×10^{7}	0.755
W + Ss + Ge + Poc	0.013	5.068×10^{-8}	4.004×10^{-6}	1.497×10^{7}	0.763
W + Ss + Poc	0.013	4.817×10^{-8}	3.806×10^{-6}	1.575×10^{7}	0.754
Le + Ss + Poc	0.013	3.788×10^{-8}	2.992×10^{-6}	2.003×10^{7}	0.753
W + Ss + Ge	0.013	2.468×10^{-8}	1.949×10^{-6}	3.074×10^{7}	0.752
Le + Ss + Ge	0.013	2.443×10^{-8}	1.930×10^{-6}	3.105×10^{7}	0.752
W + F + Ge + Poc	0.013	1.226×10^{-8}	9.687×10^{-7}	6.186×10^{7}	0.758
W + Le + Ss	0.013	9.055×10^{-9}	7.153×10^{-7}	8.378×10^{7}	0.748
W + Ss	0.013	9.655×10^{-10}	7.628×10^{-8}	7.857×10^{8}	0.730
Le + Ss	0.013	3.475×10^{-10}	2.745×10^{-8}	2.183×10^{9}	0.726
W + Le + Ge	0.013	7.183×10^{-11}	5.674×10^{-9}	1.056×10^{10}	0.730
W + Le + Ge + Poc	0.013	6.835×10^{-11}	5.399×10^{-9}	1.110×10^{10}	0.739
W + Le + Poc	0.013	3.995×10^{-11}	3.156×10^{-9}	1.899×10^{10}	0.727
Le + F + Ge	0.013	3.838×10^{-11}	3.032×10^{-9}	1.977×10^{10}	0.727
Le + F	0.013	2.344×10^{-11}	1.852×10^{-9}	3.236×10^{10}	0.715
Le + F + Poc	0.013	8.976×10^{-12}	7.091×10^{-10}	8.452×10^{10}	0.721
Le + F + Ge + Poc	0.013	6.562×10^{-12}	5.184×10^{-10}	1.156×10^{11}	0.729
W + Ge	0.013	4.111×10^{-12}	3.248×10^{-10}	1.845×10^{11}	0.708
W + Ge + Poc	0.013	3.515×10^{-12}	2.777×10^{-10}	2.158×10^{11}	0.717
W + Le	0.013	2.243×10^{-12}	1.772×10^{-10}	3.382×10^{11}	0.705
W + Poc	0.013	1.730×10^{-12}	1.366×10^{-10}	4.386×10^{11}	0.704
W	0.013	9.747×10^{-14}	7.701×10^{-12}	7.782×10^{12}	0.679
Le + Ge + Poc	0.013	1.394×10^{-14}	1.101×10^{-12}	5.442×10^{13}	0.693
Le + Ge	0.013	1.326×10^{-14}	1.048×10^{-12}	5.719×10^{13}	0.682
Ss + F + Ge	0.013	3.208×10^{-15}	2.534×10^{-13}	2.365×10^{14}	0.687
Ss + F + Ge + Poc	0.013	2.238×10^{-15}	1.768×10^{-13}	3.389×10^{14}	0.695
Le + Poc	0.013	1.655×10^{-15}	1.307×10^{-13}	4.584×10^{14}	0.672
Ss + F + Poc	0.013	7.308×10^{-16}	5.774×10^{-14}	1.038×10^{15}	0.680
Ss + Ge + Poc	0.013	2.201×10^{-16}	1.739×10^{-14}	3.446×10^{15}	0.674
Ss + F	0.013	1.144×10^{-16}	9.036×10^{-15}	6.632×10^{15}	0.659
Ss + Ge	0.013	3.897×10^{-17}	3.079×10^{-15}	1.947×10^{16}	0.654
Le	0.013	3.039×10^{-17}	2.400×10^{-15}	2.497×10^{16}	0.639
Ss + Poc	0.013	9.778×10^{-18}	7.725×10^{-16}	7.758×10^{16}	0.647
Ss	0.013	4.226×10^{-21}	3.339×10^{-19}	1.795×10^{20}	0.590
F + Ge	0.013	1.565×10^{-32}	1.237×10^{-30}	4.846×10^{31}	0.417
F + Ge + Poc	0.013	5.599×10^{-33}	4.424×10^{-31}	1.355×10^{32}	0.424
Ge + Poc	0.013	6.897×10^{-36}	5.448×10^{-34}	1.100×10^{35}	0.346
F + Poc	0.013	6.589×10^{-36}	5.206×10^{-34}	1.151×10^{35}	0.345
Ge	0.013	1.971×10^{-36}	1.557×10^{-34}	3.849×10^{35}	0.313
F	0.013	9.958×10^{-37}	7.867×10^{-35}	7.618×10^{35}	0.306
Poc	0.013	2.300×10^{-41}	1.817×10^{-39}	3.298×10^{40}	0.188
Null model	0.013	1.435×10^{-46}	1.134×10^{-44}	5.286×10^{45}	0.000

G.2 Residuals versus Predictions for log-Wealth

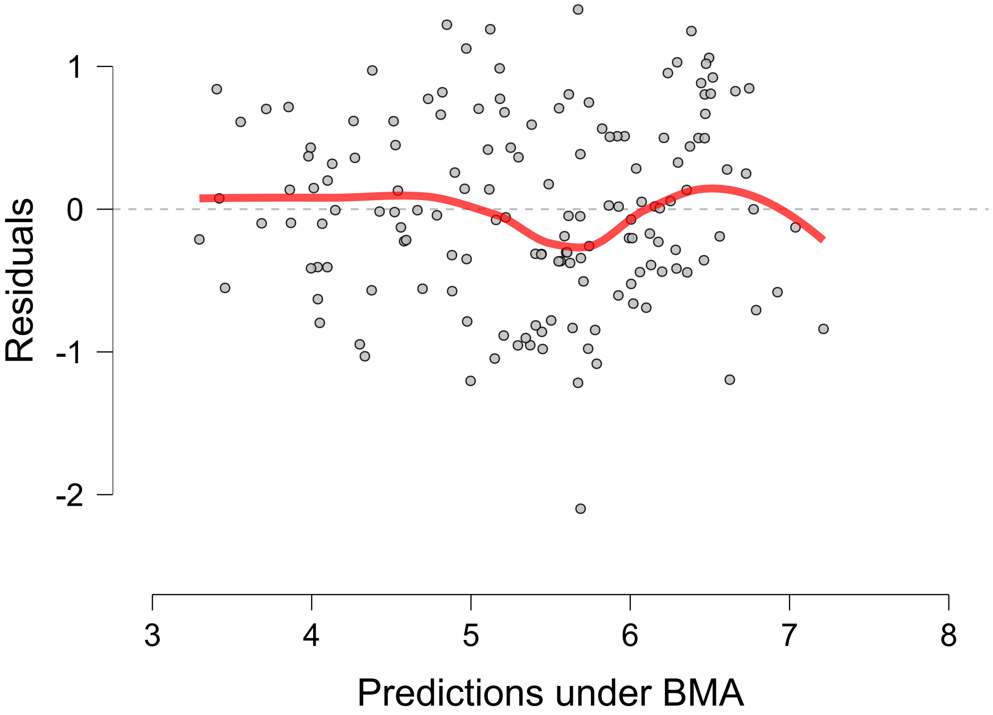


Figure G.1: Assumptions checks for Happiness predicted by log-transformed Wealth. In contrast to the right panel of Figure 8.2, the red line is completely flat and the variance is approximately constant across the predicted values.

H
Appendix of Chapter 9

H.1 Parameter Estimates for the Sorting Hat Data

Table H.1: Parameter estimates for each of the houses in the data of Jakob et al. (2019). The interpretation of each column is identical to that of Table 9.4.

				95% CI	
Predictor	Level	Mean	SD	Lower	Upper
Intercept		26.923	0.215	26.46	27.337
Sorting house	Gryffindor	−0.568	0.357	−1.28	0.140
	Hufflepuff	−2.610	0.360	−3.34	−1.898
	Ravenclaw	−0.696	0.330	−1.36	−0.037
	Slytherin	3.874	0.418	3.04	4.719

I

Appendix of Chapter 10

I.1 Within-Participant Effects with Sphericity Corrections

Table I.1: Within-Participant Effects with Sphericity Corrections. Identical to Table 10.1 but includes sphericity corrections. The general results remain unchanged. Table from JASP.

Cases	Sphericity Correction	Sum of Squares	df	Mean Square	F	p
PT	—	57532.64	1.00	57532.64	2.24	.152
Residuals	—	461800.53	18.00	25655.59		
Congruency	—	33727.79	2.00	16863.89	22.16	< .001
	Greenhouse-Geisser	33727.79	1.51	22314.52	22.16	< .001
	Huynh-Feldt	33727.79	1.62	20814.49	22.16	< .001
Residuals	—	27398.21	36.00	761.06		
	Greenhouse-Geisser	27398.21	27.21	1007.05		
	Huynh-Feldt	27398.21	29.17	939.35		
PT * Congruency	—	3124.49	2.00	1562.25	2.29	.115
	Greenhouse-Geisser	3124.49	1.33	2345.32	2.29	.137
	Huynh-Feldt	3124.49	1.39	2233.94	2.29	.134
Residuals	—	24488.84	36.00	680.25		
	Greenhouse-Geisser	24488.84	23.98	1021.22		
	Huynh-Feldt	24488.842	25.18	972.72		

I.2 Changes to Bayesian ANOVAs in JASP

The latest version of JASP (0.16.3) introduces additional changes designed to increase the flexibility of Bayesian ANOVA, which we discuss below. In contrast to the modified model specification presented in the main text, these additional changes have no consequence for the results of previous analyses.

I.2.1 Principle of Marginality

A long-standing debate in the statistical literature concerns which models to compare when testing main effects in the presence of interactions.[1] One option is to compare the complete model, containing all possible main effects and interactions, to the nested model that omits the to-be-tested main effect. In our example, the model `PT + PT * Congruency` would be compared to `PT + Congruency + PT * Congruency`. This top-down approach is recommended by the US Food and Drug Administration (1988) and corresponds to Type III-sums of squares in frequentist ANOVA. Proponents of the principle of marginality reject the top-down approach (Nelder, 1977; Venables, 2000): They argue that testing main effects in the presence of interactions, while possible, tests practically nonsensical hypotheses (p. 50 Nelder, 1977). Therefore, analysts should proceed to test simple effects rather than main effects. Accordingly, the principle of marginality demands that a model which includes an interaction must include all main effects that are marginal to (i.e. part of) it. The top-down model comparison violates the principle of marginality because the null model `PT + PT * Congruency` omits the main effect `PT` that is marginal to the interaction `PT * Congruency`. A test of main effects that respects the principle of marginality compares a model including only main effects to the nested model that omits the to-be-tested main effect. In our example, the model `PT` would be compared to `PT + Congruency`. This approach corresponds to Type II-sums of squares in frequentist ANOVA.

Because the principle of marginality is a general statement about model specification, the controversy is not limited to pairwise model comparisons. In a model averaging context, proponents of the principle of marginality argue to exclude all models that violate the principle from consideration (i.e. assign a prior probability of 0), rather than considering every possible model (Rouder et al., 2016).

It is worth noting that the two approaches only diverge if the effects are correlated, i.e. main and interaction effects compete to account for variance in the dependent variable. Effects may be correlated when, for example, independent variables are observed rather than manipulated or when the design is unbalanced design. If all effects are uncorrelated, both approaches will yield the same results.

For frequentist ANOVA, JASP users can choose either Type II or Type III sums of squares (with the latter, violating the principle of marginality, being

[1] The same considerations apply to testing lower-order (e.g., two-way) interactions in the presence of a higher-order (e.g., three-way) interaction.

the default). In contrast, the Bayesian ANOVA in JASP previously enforced the principle of marginality, both in pairwise model comparisons and model averaging. Now, JASP also allows Bayesians to consider the complete model space and perform the pairwise model comparisons recommended by the Food and Drug Administration. As is customary, in repeated-measures ANOVA the principle of marginality is only applied to fixed effects; we include all random slope effects in all models.[2] Whether the principle of marginality should extend to random slopes is subject of current debate (Heathcote & Matzke, 2021; Rouder et al., 2022; van Doorn et al., 2021).

I.2.2 Model priors

A change to all Bayesian ANOVAs is that the prior over the models can be adjusted. Previously, we used a uniform model prior by default. This means that the prior probability of each model is equal to one divided by the total number of models. However, the uniform model prior does not penalize for model complexity and a-priori favors models with half of the total predictors. We now provide five alternatives, the Beta-binomial prior (Scott & Berger, 2010), the Wilson prior (M. A. Wilson et al., 2010), the Castillo prior (Castillo et al., 2015), the Bernoulli prior, and a custom option. The Beta-binomial prior assigns prior mass to the number of included predictors and then distributes this mass equally across all models with that number of predictors. For example, given a Beta-binomial prior (1, 1) the prior probability of the set of models that include one predictor is equal to the set of models with two predictors. The prior probability of a specific model that includes one predictor can be obtained by dividing the prior probability of including one predictor by the number of models that include one predictor. The Wilson prior and Castillo prior are variants of the Beta-binomial prior tailored to large designs with many predictors. The Bernoulli prior requires the specification of a prior probability p for including any predictor. If the total number of variables is denoted K and a particular model includes j variables then the prior probability of that model is given by $p^j(1-p)^{K-j}$. A straightforward extension of the Bernoulli prior is to specify a value for p individually for each predictor, which is the manual prior.

I.2.3 Parameter priors

Aside from prior distributions over models, the Bayesian ANOVA also requires the specification of a prior distribution on the effects within a model (i.e.,

[2] In our example, the unabridged specification including random slopes reads `PT + participant + participant * PT + participant * Congruency` and `PT + Congruency + participant + participant * PT + participant * Congruency`.

the coefficients). Following Rouder et al. (2012), we use the Jeffreys–Zellner–Siow prior. This prior has one hyperparameter called r, which determines the width of this distribution. Previously, one value of r could be specified for the groups of fixed effects, covariates, and random effects. Now, analysts have more flexibility: It is possible to supply separate values of r for any individual fixed and random effects considered.

I.3 Model Table for Simultaneous Fixed and Random Effects

Table I.2: Bayesian Comparisons of Models while introducing Fixed and Random Effects Simultaneously. Model formulas simultaneously introduce fixed effects and random slopes (i.e. `PT + PT * participant`). P(M) and P(M|data), respectively, indicate prior and posterior model probabilities; BF10 indicates Bayes factors relative to the best performing model; error is the relative error associated with the numerical method used to estimate the Bayes factors.

Models	P(M)	P(M|data)	BF10	error
PT + Congruency	0.200	0.615	1.000	
PT + Congruency + PT * Congruency	0.200	0.385	0.627	3.893
PT	0.200	$2.235 \cdot 10^{-5}$	$3.636 \cdot 10^{-5}$	9.973
Null model (incl. subject and random slopes)	0.200	$1.637 \cdot 10^{-26}$	$2.663 \cdot 10^{-26}$	1.903
Congruency	0.200	$7.185 \cdot 10^{-30}$	$1.169 \cdot 10^{-29}$	1.906