Bayesian Sample Size Determination for Longitudinal Intervention Studies with Linear and Log-linear Growth

Ulrich Lösener, Mirjam Moerbeek, and Herbert Hoijtink

Utrecht University

Department of Methodology and Statistics

Author Note

Ulrich C. Lösener,

u.c.losener1@uu.nl

This project was funded by the NWO Open Competition Grant (406.21.GO.006).

There are no conflicts of interest to declare.

Abstract

A priori sample size determination (SSD) is essential in designing cost-efficient trials and in avoiding underpowered studies. In addition, reporting a solid justification for a certain sample size is required by most ethical committees and many funding agencies. Most often SSD is based on null hypothesis significance testing (NHST), an approach that has received severe criticism in the past decades. As an alternative, Bayesian evaluation of informative hypotheses has been developed. Informative hypotheses reflect specific theoretical and/or empirical expectations using (in)equality constraints on model parameters. Bayes factors quantify the relative support in the data for informative hypotheses (including the null hypothesis) without suffering from some of the drawbacks of NHST. SSD for Bayesian hypothesis testing relies on simulations and has only been studied recently. Available software for this is limited to simple models such as ANOVA and t-test, in which observations are assumed to be independent from each other. However, this assumption is rendered untenable when employing a longitudinal design where observations are nested within individuals. In that case, a multilevel model should be used. This paper provides researchers with an invaluable tool to perform SSD for multilevel models with longitudinal data in a Bayesian framework along with the necessary theoretical background and concrete empirical examples. The open source R function that enables researchers to tailor the simulation to their trial at hand can be found on the GitHub page page of the first author.

Keywords: Sample Size Determination, Sample Size Estimation, Multilevel Model, Bayes Factor, Power, Monte Carlo Simulation, Approximate Adjusted Fractional Bayes Factor, Linear Growth, Log-linear Growth

Introduction

Statistical power is defined as the probability of finding an existing effect in the data at hand. The power of an experiment depends on the sample size and the effect size as well as the probability of making a type I error (false positive; Cohen, 2013). Researchers strive for high power in order to have high chances of finding an existing effect. Determining the minimal sample size necessary to achieve a certain power level is therefore a vital component of the planning phase in an experiment. This procedure is referred to as "sample size determination" (SSD) and it is required by most ethical committees and many funding agencies. The benefit of SSD is two-fold: On the one hand, it serves to avoid underpowered studies where the hypothesized effect may exist in the population but the researcher fails to detect it because their sample size is too small. This inferential error of falsely concluding that there is no effect is called type II error, and underpowered studies remain a major problem in psychological science (Maxwell, 2004; Vadillo, Konstantinidis, & Shanks, 2016). On the other hand, SSD can avoid overpowered studies with huge sample sizes where even small differences become statistically significant, potentially leading to flawed and overconfident conclusions (Faber & Fonseca, 2014; Kaplan, Chambers, & Glasgow, 2014). In both cases of inadequate sample size, resources are wasted and unnecessary strain is put on participants (Case & Ambrosius, 2007), resulting in unethical research practice.

Following widespread criticism of the frequentist approach to null hypothesis significance testing (NHST) using *p*-values, Bayesian hypothesis evaluation employing the Bayes Factor has been developed as an alternative inferential tool (Jeffreys, 1935; Kass & Raftery, 1995). The use of BFs is rapidly gaining popularity among researchers (Schmalz, Biurrun Manresa, & Zhang, 2023; Van De Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Despite this, the algorithms available for SSD within the Bayesian framework are currently limited to simple models such as ANOVA and t-test. However, in psychological research, statistical models are often more sophisticated. For example, in the presence of longitudinal data, multilevel regression models are the method of choice (Hedeker & Gibbons, 2006). Using multilevel models, we can evaluate the effectiveness of a treatment intervention over time by comparing the mean growth trajectories of an experimental and control condition. The power of such experiments also depends on the number of measurements and their location in time, which is why we cannot simply use the SSD results from a t-test or an ANOVA. To our knowledge, there is no available software to perform Bayesian SSD on these types of more complex models yet. In this paper, we address the scarcity of available software for Bayesian SSD for trials with longitudinal data with linear or log-linear growth by introducing an open-access R function available on GitHub. This function performs Bayesian SSD for multilevel models using Monte-Carlo simulation. For now, we focus on the case where two competing hypotheses are formulated about linear or log-linear growth. This aims at extending the work by Fu et al. (2021) to the more complex case of multilevel models with linear as well as log-linear growth.

The subsequent sections of this paper are organized as follows. First, we describe the mutilevel model which is used in this paper. Next, we elaborate on the Bayes Factor generally and the Approximate Adjusted Bayes Factor specifically, followed by a brief comment on power in a Bayesian context. Subsequently, we illustrate our method using two empirical examples, followed by a description of the algorithm employed in our simulation. Finally, we provide a summary of our study results as well as directions for future research.

The Multilevel Model for Longitudinal Data

Suppose we want to assess the effectiveness of a new psychological treatment intervention on life satisfaction in an experimental setting with randomized allocation to either the treatment or the control condition. To this end, each individual's life satisfaction scores are measured repeatedly over time, meaning that we are dealing with longitudinal data. This is a typical situation in psychological intervention research, for which multilevel models have been implemented for a long time (Bryk & Raudenbush, 1987). The reason why many traditional methods such as ANOVAs and single-level regression are not suitable for this type of data is that the assumption of independence of observations is not tenable (Walsh, 1947; Raudenbush & Bryk, 2002). This is because observations belonging to the same individual tend to be more similar than observations across individuals. Employing models that ignore the nesting of observations within individuals may lead to type I or type II error inflation (Moerbeek, van Breukelen, & Berger, 2003). Alternatively, repeated measures ANOVA can be used, but this model relies on strict assumptions such as compound symmetry (constant covariances over timepoints) and equal spacing of time points (Hox, Moerbeek, & Van de Schoot, 2018). Furthermore, with repeated measures ANOVA, all measurements of an individual need to be excluded from the analysis if there is a missing value on just one occasion. This results in a huge decrease of power when attrition is present. However, these drawbacks do not apply to multilevel models in which the nested structure of the data is explicitly modeled using a hierarchical set of regression equations (Raudenbush & Bryk, 2002; Goldstein & Browne, 2003; Bosker & Snijders, 2011). Also, multilevel models allow for the measurements to be irregularly spaced in time. For a comprehensive overview of multilevel models, see the book by Hox et al. (2018). In this paper, we will limit our scope to modeling linear and log-linear growth as they represent popular choices in psychological intervention research (e.g., Althammer, Reis, Van der Beek, Beck, & Michel, 2021). In linear growth models, it is assumed that changes occur at the same rate over time while log-linear models assume that the rate of change becomes smaller over time. Also, we consider the case where measurement occasions are not equally spaced, meaning that the frequency of observation differs over time which is quite common in longitudinal clinical trials (Gibbons, Hedeker, & DuToit, 2010). The regression equation for the first (measurement occasion) level is

$$Y_{ij} = \pi_{0i} + \pi_{1i}T_j + e_{ij} \qquad \text{with } e_{ij} \sim N(0, \sigma^2), \tag{1}$$

where Y_{ij} the level of life satisfaction for individual $i = 1, ..., N$ at measurement occasion $j = 1, ..., n$, π_{0i} is the intercept, π_{1i} is the regression coefficient of the time variable T_j, and e_{ij} denotes the residual. The regression equations for the second (subject) level are

$$\pi_{0i} = \beta_0 + u_{0i} \tag{2}$$

$$\pi_{1i} = \beta_1 + \beta_2 C_i + u_{1i}, \tag{3}$$

where u_{0i} is the individual deviation from the overall intercept, and u_{1i} is the individual deviation from the slope in their treatment condition. Note that the intercept π_{0i} and slope π_{1i} in Equation (1) can vary per person (hence the subscript i) and it is expected that some of the slope variance can be explained by the binary predictor "treatment condition" (C_i). The average intercept is denoted by β_0 and the average coefficient π_{1i} for the control group $(C_i = 0)$ is denoted by β_1. The coefficient β_2 indicates how much the average slope differs between treatment groups. The population distribution of the random effects $\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix}$ is assumed to be bivariate normal with means of zero $N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u)$ with the variance-covariance matrix

$$\Sigma_u = \begin{pmatrix} \sigma^2_{u0} & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma^2_{u1} \end{pmatrix}.$$

By substituting Equation (2) and (3) into Equation (1) we obtain the combined regression equation containing both levels:

$$Y_{ij} = \beta_0 + u_{0i} + (\beta_1 + \beta_2 C_i + u_{1i}) * T_j + e_{ij} \tag{4}$$

or, when rearranging the terms:

$$Y_{ij} = \beta_0 + \beta_1 T_j + \beta_2 C_i T_j + u_{0i} + u_{1i} T_j + e_{ij} \tag{5}$$

The parameter of interest in this model when examining a potential treatment effect is β_2, indicating the magnitude of interaction between time (T_j) and treatment condition (C_i). Thus, β_2 represents the differential growth rate of life satisfaction across time in the two experimental conditions. If $\beta_2 = 0$, then individuals in the treatment and control condition exhibit the same growth/decline in terms of life satisfaction, implying that no treatment effect is present. If $\beta_2 > 0$, then individuals in the treatment condition exhibit a larger increase of life satisfaction, implying that there is a positive treatment effect. Two typical hypotheses on a potential intervention effect can be formulated as follows.

$$\mathcal{H}_0 : \beta_2 = 0 \tag{6}$$

$$\mathcal{H}_1 : \beta_2 > 0 \tag{7}$$

As shown by Moerbeek and Teerenstra (2015), the standard error of the estimated β_2 is defined as

$$\sigma_{\hat{\beta}_2} = \sqrt{\frac{4(\sigma_e^2 + \sigma_{u1}^2 \sum_{j=0}^{n-1} T_j^2)}{N \sum_{j=0}^{n-1} T_j^2}}, \tag{8}$$

where T_j is the point in time of measurement j and n is the number of measurement occasions. The fact that the level 2 sample size (number of individuals, N) is in the denominator of Equation (8) indicates that $\sigma_{\hat{\beta}_2}$ decreases with larger samples. Therefore, our estimation of β_2 becomes increasingly precise as more individuals are taken into account. This is rather intuitive yet crucial for the logic of SSD: We need to increase N until our estimation of β_2 is sufficiently precise to make a sound inferential decision. Within the frequentist framework, SSD can be done analytically via closed-form equations relating $\sigma_{\hat{\beta}_2}$ to statistical power (Moerbeek & Teerenstra, 2015). In this paper, however, we employ the BF for which there are no closed-form solutions for SSD and we therefore rely on simulation (see Fu, 2022).

Effective sample size in in multilevel models

An important issue in Bayesian hypothesis evaluation for multilevel models is the quantification of the total sample size which is often referred to as the "effective sample size", N_{eff} (Berger & Pericchi, 1996). The assumption that $N_{eff} = N * n$ (where N and

n are the sample sizes of level 2 and 1, respectively) is too optimistic, as measurements within individuals are correlated with each other and can therefore not be counted as independent observations. The effective sample size is therefore somewhere between $N * n$ and N, depending on the degree of correlation of observations within individuals. The degree of correlation is expressed in the *Intraclass Correlation Coefficient* (ICC; Killip, Mahfoud, & Pearce, 2004) which can only be obtained for the intercept-only model. However, because we are using a multilevel model with random slopes, we cannot calculate N_{eff} via the ICC (Sekulovski & Hoijtink, 2023). Sekulovski and Hoijtink (2023) propose an alternative method to derive N_{eff} via multiple imputation techniques (Van Buuren, 2018), however, this approach is more suitable for cross-sectional applications and is too computationally intensive for our simulation study. We therefore conservatively assume the "worst case scenario" where $N_{eff} = N$. This also means that the sample size recommendation provided in our function is the upper bound to achieve the desired power, and the actual power level might be slightly higher. [1]

Study duration and frequency of observation

In our model, the measurement occasions $j = 1, ..., n$ are assumed to be the same for each individual (as indicated by the absence of the subscript i in T_j). However, the measurement occasions do not need to be equally spaced in time. This means that the model allows, for example, for more frequent measurements at the beginning and end of an experiment versus halfway through a study. The duration of a study is denoted by D. In the case of equally spaced measurements, the frequency of observation f represents the number of measurements taken in each unit time. Assuming that the first measurement is taken at baseline (i.e., time point zero), one can derive the time vector T_j using D and f as follows: $T_j = ((j-1)/f)_{j=1}^{D}$. Hence, a study with duration D and frequency of observation f has a total of $n = fD + 1$ measurement occasions and the study terminates at time $D = (n-1)/f$ (Raudenbush & Liu, 2001). In case measurements are spaced irregularly, f is the average number of measurements per unit time. One of our interests

[1] Note that, in case there is good reason to believe that N_{eff} is closer to $N * n$ than it is to N, it is possible to set $N_{eff} = N * n$ (i.e., the best case scenario) via the `Neff` argument of the function (see Table 1).

lies in investigating the effect of f and D on the power of a longitudinal intervention study.

The tool we employ to evaluate $\mathcal{H}_0$ and $\mathcal{H}_1$ is the Approximate Adjusted Fractional Bayes Factor (AAFBF; Gu, Mulder, & Hoijtink, 2018). The following section offers an introduction to Bayes Factors in general and the AAFBF in particular, along with a brief subsection on informative hypotheses.

The Bayes Factor

In this section, we provide a general overview of the Bayes Factor (BF) and some of its advantages over traditional null hypothesis significance testing (NHST) methods. Subsequently, we briefly introduce the concept of informative hypotheses before elaborating on the specific way of computing the BF in this paper.

The BF (Kass & Raftery, 1995) quantifies the relative support in the data for a pair of competing hypotheses. It is defined as the ratio of the marginal likelihoods under the two hypotheses. Therefore,

$$BF_{01} = \frac{m(X \mid \mathcal{H}_0)}{m(X \mid \mathcal{H}_1)}, \tag{9}$$

where X is the data at hand and $m(X \mid \mathcal{H})$ denotes the marginal likelihood under $\mathcal{H}$. The prior odds reflect the probability assigned to a hypothesis relative to the probability of the competing hypothesis *before* considering any data. If $\mathcal{H}_0$ is a priori considered to be twice as likely than $\mathcal{H}_1$, then the prior odds would be $\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} = 2$. The posterior odds $\frac{P(\mathcal{H}_0|X)}{P(\mathcal{H}_1|X)}$ denote the updated belief about the relative probability of a hypothesis being true *after* considering the data. The posterior odds are calculated by multiplying the prior odds with the BF.

$$\frac{P(\mathcal{H}_0 \mid X)}{P(\mathcal{H}_1 \mid X)} = BF_{01} * \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} \tag{10}$$

Hence, the BF reflects how our beliefs about the odds of a pair of hypotheses change after considering the data at hand (Lavine & Schervish, 1999; Bernardo & Smith, 2009). Note that in this paper, we always assume that both hypotheses are equally likely a priori such that $\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} = 1$.

The BF is the most prominent tool for hypothesis evaluation and model selection in the Bayesian framework (Kass & Raftery, 1995) and a viable alternative to p-values as it does not suffer from the pitfalls of frequentist inference (Hoijtink, Klugkist, & Boelen, 2008), some of which will be mentioned here. First, the interpretation of the BF is simple and intuitive. As a quantification of evidence for a certain hypothesis in comparison to another, $BF_{01} = 5$ means that the data supports $\mathcal{H}_0$ five times more than $\mathcal{H}_1$. Conversely, $BF_{01} = 0.2$ means that the data supports $\mathcal{H}_1$ five times more than $\mathcal{H}_0$. Contrarily, widespread misinterpretations of p-values persist in social and psychological science, resulting in incorrect or flawed inferences (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). Second, BFs can provide evidence *in favor* of the null hypothesis (Hoijtink, Mulder, van Lissa, & Gu, 2019; Keysers, Gazzola, & Wagenmakers, 2020), while in the framework of NHST, $\mathcal{H}_0$ can only be rejected but never accepted. Third, with increasingly large samples, hypothesis tests employing BFs do not become biased towards rejection of $\mathcal{H}_0$ (Hoijtink, van Kooten, & Hulsker, 2016) as is the case in NHST (Lantz, 2013). Instead, they simply reflect the cumulative evidence for the best hypothesis under consideration, increasingly favoring it as additional data are analyzed. Fourth, as shown in Equation (10), the probability of a hypothesis being true given the data $P(\mathcal{H} \mid X)$ can be calculated. Contrarily, in NHST one can only compute the probability of finding the test statistic (which is a function of X) or a more extreme one, given that $\mathcal{H}_0$ is true, $P(X \geq x \mid \mathcal{H}_0)$. Researchers typically strive to learn about the former quantity rather than the latter because they are interested in the probability of their hypothesis being correct rather than the probability of finding their results (or more extreme ones) given $\mathcal{H}_0$ (Gill, 1999). Simulation studies have shown that in many scenarios, the correlation between these two conditional probabilities is quite low (Trafimow & Rice, 2009). This finding suggests that knowing $P(X \geq x \mid \mathcal{H}_0)$ does not necessarily give insight in $P(\mathcal{H} \mid X)$. The only scenario where $P(\mathcal{H} \mid X \geq x) = P(X \geq x \mid \mathcal{H})$ is when $P(\mathcal{H}) = P(X \geq x)$, and there is usually no justification to assume this equality (Gill, 1999). Worse still, many researchers are convinced to have obtained $P(\mathcal{H} \mid X)$ after inspecting p-values, resulting in incorrect interpretations of results and inferential errors (Meehl, 1990; Hubbard, 2011). Finally, as BFs do not rely on controlling type I error rates, they can be recomputed as more data are being analyzed without the need for correction, a procedure which is referred to as "Bayesian Sequential Design" or "Bayesian Updating" (Moerbeek, 2023). One could even argue that this method can be used *instead of* of SSD because one can just stop collecting more data once a certain BF is achieved (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017; Heck et al., 2022). However, in the case of longitudinal designs as well as small population sizes (such as in rare diseases), updating is not suitable (Fu et al., 2021). This is because, especially when the duration of the intervention is long, adding a new wave of participants to the sample would require lots of additional resources and might increase the duration of the study significantly. Additionally, ethical committees and funding agencies often require an a priori justification of sample size which can only be done via SSD.

Informative hypotheses

In recent years, Bayesian inference using informative hypotheses has increasingly gained attention in statistical literature for its wide and straightforward applicability (Schmalz et al., 2023; Van De Schoot et al., 2017). Informative hypotheses reflect the researcher's specific expectations with respect to one or more model parameters. These expectations can be based on expert knowledge, previous findings in the literature, or subjective beliefs (Moerbeek, 2023). The way informative hypotheses express theoretical expectations is by imposing certain constraints on model parameters (Hoijtink, 2011). For example, the effect of multiple experimental conditions such as a waiting list (WL), treatment as usual (TAU), and the treatment intervention (INT) may be evaluated. If it is expected that all three group effects are equal, the resulting informative hypothesis is $\mathcal{H}_0 : \theta_{WL} = \theta_{TAU} = \theta_{INT}$. The relations between parameters can be expressed by equality constraints such as above but also of inequality constraints $(<, >)$, about equality constraints $(\approx)$, or all of the above. For example, the expectation that the INT performs better than TAU which, in turn, performs better than WL is captured by the informative hypothesis $\mathcal{H}_1 : \theta_{WL} < \theta_{TAU} < \theta_{INT}$. When formulating about equality constraints, it is hypothesized that a difference between parameters does not exceed a certain threshold, for example, a minimal clinically relevant difference ($\mathcal{H}_2 : \theta_{TAU} \approx \theta_{INT}$) or equivalently, $\mathcal{H}_3 : \theta_{TAU} - \theta_{INT} < threshold$). It is furthermore possible to not constrain certain parameters in any way. For example, the expectation that the waiting list performs worse than the other two conditions while there is no expectation about the ordering between treatment as usual and the intervention can be expressed as $\mathcal{H}_4 : \theta_{WL} < (\theta_{TAU}, \theta_{INT})$. Informative hypothesis can be tested against each other or against the unconstrained hypothesis $\mathcal{H}_u : \theta_{WL}, \theta_{TAU}, \theta_{INT}$ which imposes no constraints on the parameters at all and can be looked upon as the "fail-safe" hypothesis if all other hypotheses describe the data poorly (Hoijtink, Mulder, et al., 2019). Alternatively, it is possible to test one or multiple hypotheses against their complement hypothesis $\mathcal{H}_c$. For example, the complement of $\mathcal{H}_1 : \theta_{WL} < \theta_{TAU} < \theta_{INT}$ is simply $\mathcal{H}_c$: not $\mathcal{H}_1$. The complement covers all of the parameter space which is not covered by the union of the hypotheses it is evaluated against. Note that in case the hypothesis only contains equality constraints as in $\mathcal{H}_0$, the complement is equal to the unconstrained hypothesis (Hoijtink, Mulder, et al., 2019). By virtue of this flexibility in imposing constraints on (many) model parameters, informative hypotheses cover a much broader range of expectations compared to NHST and provide one with the appealing possibility of comparing multiple informative hypotheses with each other (Gu et al., 2018; Hoijtink, Mulder, et al., 2019).

The Approximate Adjusted Fractional Bayes Factor

A number of different approaches to calculating the BF can be found in the literature but when evaluating informative hypotheses about multilevel model parameters, the *Approximate Adjusted Fractional Bayes Factor* (henceforth abbreviated as BF) stands out for its convenience and straightforward computation (Mulder, 2014; Gu et al., 2018). While Gu, Mulder, and Hoijtink (2018) provide a more extensive overview of the this way of calculating the BF, we limit ourselves to briefly describing its main characteristics.

As shown by Mulder et al., (2010), Equation (9) for the BF of a hypothesis i

against the unconstrained hypothesis $\mathcal{H}_u$ can be rewritten as

$$BF_{iu} = \frac{fit_i}{comp_i} \tag{11}$$

and the BF between two competing hypotheses $\mathcal{H}_i$ and $\mathcal{H}'_i$ can be expressed by

$$BF_{ii'} = \frac{fit_i/comp_i}{fit_{i'}/comp_{i'}}, \tag{12}$$

where fit_i is the relative fit and $comp_i$ the relative complexity of hypothesis i (Gu et al., 2018). Note that for a hypothesis with only equality constraints, expression (10) is equal to the so-called Savage-Dickey ratio (Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

The complexity of hypothesis i is defined as the proportion of the parameter space under $\mathcal{H}_i$ that is in accordance with the prior, indicating how general the predictions of $\mathcal{H}_i$ are. The more general (i.e., less specific) the prediction, the higher the complexity. Fit on the other hand indicates how well $\mathcal{H}_i$ describes the data, being defined as the proportion of the parameter space that is in accordance with the posterior. If the fit of both hypotheses is equal, the BF will prefer the least complex hypothesis, acting as an Occam's Razor (Hoijtink, Mulder, et al., 2019).

In this particular way of computing the BF, a fraction of the data (X^b) is used to construct a default prior. This means that the researcher does not need to specify the distributional form of the prior (O'Hagan, 1995) and the subjective input from the researcher is only expressed in the form of the specified hypotheses and the choice of the size of X^b. This addresses a common critique stating that BFs are too sensitive to the prior specified by the researcher, and applied researchers are often uncertain about how to construct said prior (Frick, 1996; Trafimow, 2003).

Furthermore, a normal distribution is used to approximate the prior and the t-distributed posterior. The prior distribution is adjusted such that it is centered around the boundary of the parameter space constrained by the hypotheses at hand. In both our examples of the psychological treatment intervention studies with hypotheses $\mathcal{H}_0$ (6) and

$\mathcal{H}_1$ (7), the focal point and thus the center of the prior is zero. The unconstrained prior $h_u(\beta_2 \mid X^b)$ and posterior $g_u(\beta_2 \mid X)$ are then approximated by

$$h_u(\beta_2 \mid X^b) \approx N(0, \sigma^2_{\hat{\beta}_2}/b) \tag{13}$$

and

$$g_u(\beta_2 \mid X) \approx N(\hat{\beta}_2, \sigma^2_{\hat{\beta}_2}), \tag{14}$$

respectively. The data at hand is denoted by X and the fraction of the data used to inform the prior is denoted by X^b. The maximum likelihood estimate of our parameter of interest is $\hat{\beta}_2$ and its squared standard error is $\sigma^2_{\hat{\beta}_2}$. The formulae for the fit of the equality constrained hypothesis $\mathcal{H}_0$ and of the inequality constrained hypothesis $\mathcal{H}_1$ are therefore

$$fit_0 = g_u(\beta_2 = 0 \mid X) \tag{15}$$

and

$$fit_1 = \int_{\beta_2 > 0} g_u(\beta_2 \mid X) d\beta_2, \tag{16}$$

respectively. Finally, the complexities of hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ are defined by

$$comp_0 = h_u(\beta_2 = 0 \mid X^b) \tag{17}$$

and

$$comp_1 = \int_{\beta_2 > 0} h_u(\beta_2 \mid X^b) d\beta_2, \tag{18}$$

respectively. Thus, the BFs for $\mathcal{H}_0$ and $\mathcal{H}_1$ as formulated in (6) and (7) against the unconstrained hypothesis $\mathcal{H}_u$ can be expressed analytically as

$$BF_{0u} = \frac{g_u(\beta_2 = 0 \mid X)}{h_u(\beta_2 = 0 \mid X^b)} \tag{19}$$

and

$$BF_{1u} = \frac{\int_{\beta_2 > 0} g_u(\beta_2 \mid X) d\beta_2}{\int_{\beta_2 > 0} h_u(\beta_2 \mid X^b) d\beta_2}. \tag{20}$$

As mentioned before, the fraction b determines the amount of data used to specify the default prior. A typical choice for this quantity is the minimal

$$b_{min} = \frac{J}{N}, \tag{21}$$

where J is the number of independent constraints in the hypotheses under investigation and N is the effective sample size (Hoijtink, Gu, & Mulder, 2019). In our case $J = 1$ because we constrain only one model parameter. Hence, the minimal number of observations necessary to identify the parameter(s) of interest is used for the prior. This method is inspired by the principle of the minimal training sample (O'Hagan, 1995; Berger & Pericchi, 1996) and has the advantage that the amount of data used to construct the prior is minimized while the part of the data used to calculate the BF is maximized[2]. An alternative choice for b is

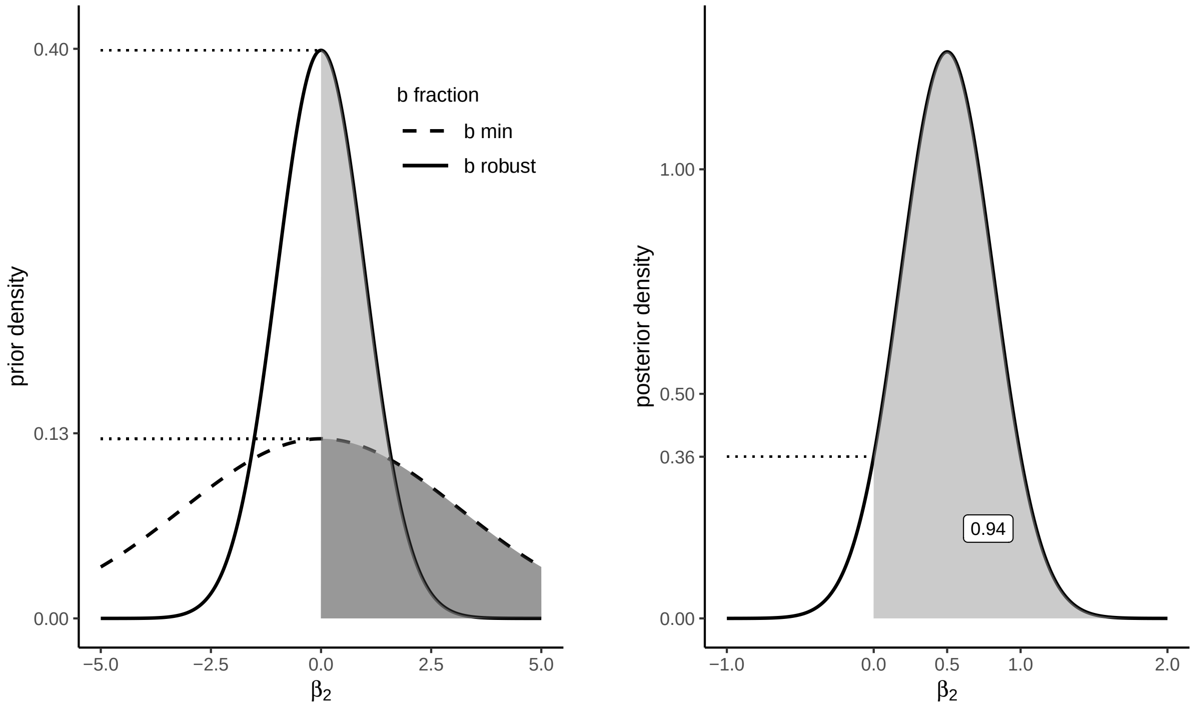
$$b_{robust} = \max\{\frac{(J+1)}{N}, \frac{1}{\sqrt{N}}\}, \tag{22}$$

generally resulting in a larger b and reducing the sensitivity of the BF to the prior distribution (O'Hagan, 1995; Conigliani & O'Hagan, 2000).

The choice of b can have non-trivial consequences of the resulting BF for hypotheses with (about) equality constraints, as will be shown in the following example.

[2] Note that two different definitions of the minimal b can be found in the literature. While Gu et al. (2018) suggest $b_{min} = \frac{J+1}{N}$, we use the b_{min} by Hoijtink et al. (2019) which is also implemented in the R-package *bain* (Gu, Hoijtink, Mulder, & van Lissa, 2021).

Suppose that after the collection of data of $N = 100$ individuals and analysis using multilevel regression, the estimate for the parameter is $\hat{\beta}_2 = 0.5$ with a corresponding squared standard error of $\sigma^2_{\hat{\beta}_2} = 0.1$. We now want to test the hypotheses (6) and (7) against each other using the BF. Figure 1 depicts the resulting complexities and fits for hypotheses (6) and (7), respectively.



(a) *Prior distributions and complexities of $\mathcal{H}_0$ and $\mathcal{H}_1$. The solid (dashed) line corresponds to the prior distribution for a relatively large (small) b. The horizontal dotted lines indicate the complexity of $\mathcal{H}_0$ under each prior (0.18 and 0.4, respectively) while the shaded area under the curve indicates the complexity of $\mathcal{H}_1$ under each prior (0.5 for both b).*

(b) *Posterior distribution for β_2 with $\hat{\beta}_2 = 0.5$ and $\sigma^2_{\hat{\beta}_2} = 0.1$. The fit for $\mathcal{H}_0$ is $g_u(\beta_2 = 0 \mid X) = 0.36$. The fit for $\mathcal{H}_1$ is $g_u(\beta_2 > 0 \mid X) = 0.94$.*

Figure 1

Prior and posterior distribution with fit and complexity for $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

The influence of b on the complexity of equality constrained hypotheses becomes apparent in Figure 1: The complexity of $\mathcal{H}_0$ increases quite drastically from .13 to .40 when using b_{robust} instead of b_{min}. Note that the complexity for the inequality constrained $\mathcal{H}_1$ stays the same because in both cases, the integral from zero to infinity is equal to .5 regardless of the prior variance. This change in complexity of $\mathcal{H}_0$ can result in substantial changes of the resulting BF: The BF of $\mathcal{H}_0$ versus the unconstrained hypothesis ($\mathcal{H}_u$) is

$BF_{0u} = 2.87$ when using b_{min} but only $BF_{0u} = 0.91$ when using b_{robust}. This phenomenon is also present when looking at the BF for $\mathcal{H}_0$ versus $\mathcal{H}_1$. The BF of $\mathcal{H}_0$ versus $\mathcal{H}_1$ is $BF_{01} = 1.52$ when using b_{min} but only $BF_{0u} = 0.48$ when using b_{robust}. This means that in these cases the BF prefers a different hypothesis over the other depending on which b is chosen. More specifically, equality constrained hypotheses such as $\mathcal{H}_0$ are more preferred by the BF when using smaller b and less preferred with larger b (Gu et al., 2018). This can be problematic because the choice of b introduces a source of subjectivity which can have an substantial influence on the BF and hence the inferential decision. However, Gu et al. (2018) provide guidelines on which b to chose in a specific situation. In our example, b_{min} would be an appropriate choice because of the small sample size ($N = 100$). This is because we want to minimize the amount of data used to construct the prior. On the other hand, when the sample size is large and robustness against a potentially misspecified prior is of importance, a larger b such as b_{robust} may be a more suitable option (O'Hagan, 1995). There are other ways to specify b, for example b_{freq} which results in equal type I and type II error probabilities (Gu, Hoijtink, & Mulder, 2016; Hoijtink, 2022), but these are beyond the scope of this article.

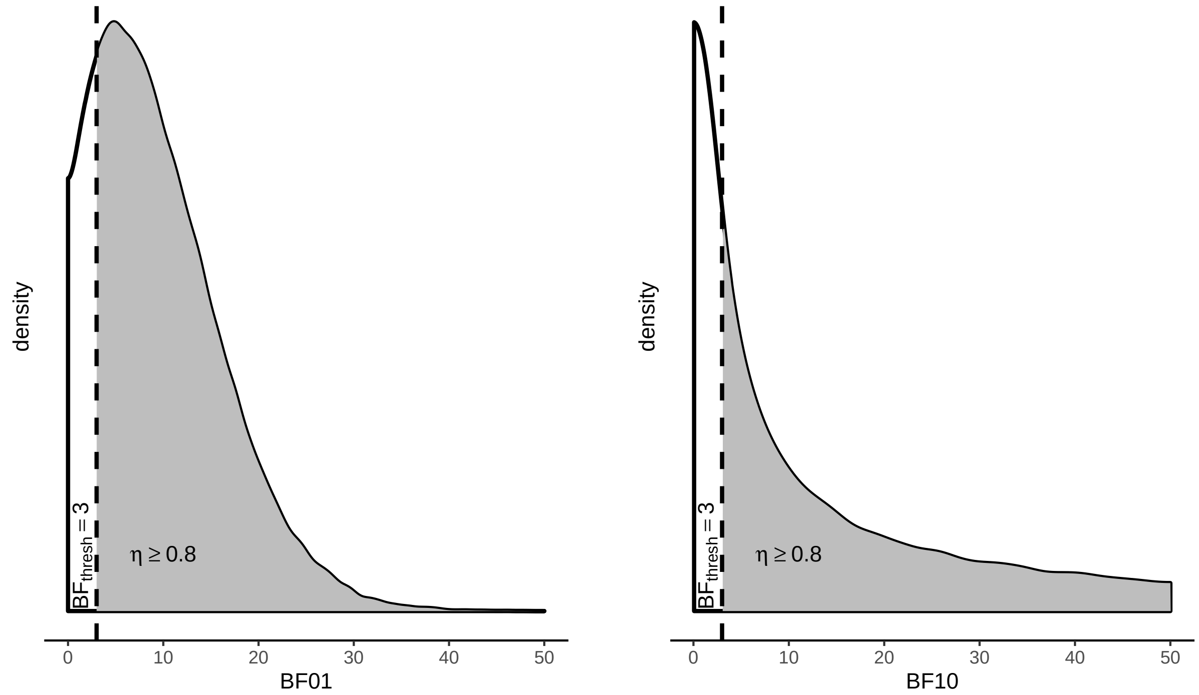
As illustrated in the above example, the BF is completely independent of b when the hypotheses at hand contain no (about-)equality constraints (Mulder, 2014). This is because the complexity (defined as the integral of the prior between the bounds of the parameter space under the hypothesis) stays the same regardless of the shape of the prior (Hoijtink, Mulder, et al., 2019). In this paper, we take the consequences of some choices for b into account by means of a sensitivity analysis. Similar to the function by Fu (2021), the result of the SSD is reported for three different choices of b when setting the argument `sensitivity = TRUE`: a) $b_{min} = \frac{1}{N}$, b) $b_2 = 2b_{min} = \frac{2}{N}$ and c) $b_3 = 3b_{min} = \frac{3}{N}$.

Bayesian Sample Size Determination

In order to perform SSD in the Bayesian framework, one first needs to specify which degree of evidence is considered compelling. This is done by defining BF_{thres}, the threshold value which the BF needs to exceed to be considered substantial. Although some authors suggest general cut-off values to interpret the strength of relative evidence indicated by the BF (Kass & Raftery, 1995), we recommend avoiding the fallacy of a global dichotomous decision rule, rather letting the relative evidence for each hypothesis speak for itself (Hoijtink, Mulder, et al., 2019). However, in the case of SSD one needs to specify some sort of decision rule in order to calculate the probability of making the correct inferential decision. We therefore let the researcher decide the amount of relative evidence that they find compelling by specifying BF_{thres} prior to executing the SSD. This threshold value may be lower (e.g., 3) for more exploratory research and higher (e.g., 10) for high-stakes research.

Next, the desired power level η needs to be established. This reflects the probability of obtaining a BF larger than BF_{thres} in favor of the true hypothesis (Fu et al., 2021). This is similar to the frequentist definition which was given previously. However, the difference is that if $\mathcal{H}_0$ is true, we also require the BF in favor of $\mathcal{H}_0$ to be at least BF_{thres} with a probability of at least η. This is a crucial contrast to frequentist power as we are dealing with two conditional probabilities instead of one. They are a) $\eta_0 = P(BF_{01} > BF_{thres} \mid \mathcal{H}_0)$, the chance of finding a BF favoring $\mathcal{H}_0$ which exceeds the threshold value, given that $\mathcal{H}_0$ is true and b) $\eta_1 = P(BF_{10} > BF_{thres} \mid \mathcal{H}_1)$, the chance of finding a BF favoring $\mathcal{H}_1$ exceeding the threshold value, given that $\mathcal{H}_1$ is true. This discrimination between η_0 and η_1 arises because in Bayesian hypothesis testing, both $\mathcal{H}_0$ and $\mathcal{H}_1$ can be accepted and are regarded as "equals" (Hoijtink, Mulder, et al., 2019; Tendeiro & Kiers, 2019). Considering that most ethics and funding committees require one single value to indicate statistical power, we by default combine the two probabilities into $\eta = min(\eta_0, \eta_1)$, where η indicates the lower bound of power for all hypotheses under consideration, similar to Wang and Gelfand, (2002). However, it is possible to perform SSD for only one of the two hypotheses via the `hyp` argument (see Table 1), reducing the necessary computing time to a minimum.

Because it is not known how to derive the sample size resulting in a certain η analytically, we resort to simulation. To do that, we need to operationalize η in a way that we can assess in the simulation. We therefore define η as the proportion of BFs favoring the true hypothesis greater than BF_{thres} across all simulated data sets. If the desired power level is, say, .80, then we look for the sample size for which at least 80% of the BFs favoring the true hypothesis exceed BF_{thres}. Figure 2 illustrates the sampling distributions of the BFs when the power criterion of $\eta = .80$ is met. It can be seen that for both hypotheses, the proportion of BFs exceeding BF_{thres} is at least .80.



(a) *Sampling distribution of all* BF_{01} *when* $\mathcal{H}_0$ *is true.*

(b) *Sampling distribution of all* BF_{10} *when* $\mathcal{H}_1$ *is true.*

Figure 2

Sampling distributions BF_{01} *and* BF_{10} *under* $\mathcal{H}_0$ *and* $\mathcal{H}_1$ *when the power criterion of* $\eta = .80$ *is met with* $BF_{thres} = 3$.

The steps of the algorithm to execute the SSD are elaborated in the following section.

Algorithm 1: Iterative increment

The function `BayeSSD` was created using R version 4.2.2 (R Core Team, 2013) and the packages "lme4" version 1.1-31 (Bates, Mächler, Bolker, & Walker, 2015) and "MASS" version 7.3-58.2 (Venables & Ripley, 2002). The logic of the function for `BayeSSD` is simple: in each iteration, m data sets with sample size N are generated under each hypothesis. Note that we assume the outcome variable to be mean centered around zero, that is, $E(Y_{ij}) = 0$. Next, the function evaluates whether the power condition is met, that is, the proportion of BFs larger than BF_{thres} favoring the true hypothesis is at least η

under both hypotheses (see Figure 2). Put differently, the power condition is met if the following inequality is true under both $\mathcal{H}_0$ and $\mathcal{H}_1$

$$\frac{\sum_{k=1}^{m} \mathbb{1}_{(BF_k > BF_{thres})}}{m} = P(BF_k > BF_{thres}) \geq \eta, \tag{23}$$

where $\mathbb{1}_{(BF_k > BF_{thres})}$ is an indicator function that is equal to 1 if the condition in brackets is met and 0 if not. The number of datasets generated under each hypothesis is denoted by m. Hence, the first and second part in expression (22) represent the proportion of BFs exceeding the threshold. If these equivalent expressions for the actual power are equal to or larger than η, then the currently evaluated N meets the power criterion. Note that BF can denote the Bayes Factor of one of the hypotheses $\mathcal{H}_0$ or $\mathcal{H}_1$ against the other one or alternatively, $\mathcal{H}_0$ or $\mathcal{H}_1$ against the complement hypothesis $\mathcal{H}_c$ or against the unconstrained hypothesis $\mathcal{H}_u$. The latter two choices are advised when comparison against a fail-safe hypothesis is appropriate (Hoijtink, Mulder, et al., 2019) and can be set via the `test` argument (see Table 1).

If this is not the case, however, we increase N by one per treatment condition until the power requirement is met. Then, the current N along with η_0 and η_1 is reported as the result of the SSD. Figure 3 illustrates each step of the procedure.

The arguments of the function along with their default values are described in Table 1.

Algorithm 2: Binary search

Because of the extensive computation time of algorithm 1 we implemented an alternative function which uses a binary search to determine the optimal sample size, reducing the number of iterations and computing time substantially. This algorithm is implemented in the final version of the function `BayeSSD`, while algorithm 1 can still be accessed on the GitHub repository of the first author. Figure 4 illustrates the procedure of a binary search algorithm. Here, the procedure is as follows (see also Fu et al., 2021): First, a minimum (N_{min}) and maximum sample size (N_{max}) are specified (e.g., 10 and 10000). Next, the power for the sample size in the exact middle of this range (N_{mid}) is

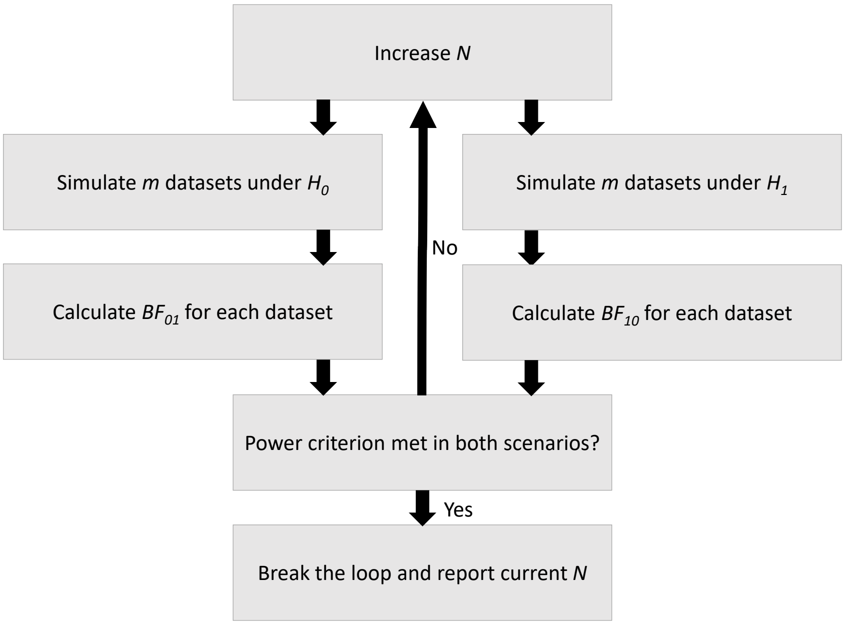

Increase *N*
Simulate *m* datasets under H_0
Simulate *m* datasets under H_1
No
Calculate BF_{01} for each dataset
Calculate BF_{10} for each dataset
Power criterion met in both scenarios?
Yes
Break the loop and report current *N*

Figure 3
Procedure of the first SSD algorithm. N = total number of participants, BF = Bayes Factor.

evaluated. If the resulting power is at least η under both hypotheses, then N_{mid} becomes N_{max} in the next iteration. If either one of the power levels is less than η, N_{mid} becomes N_{min} in the next iteration. When $N_{mid} = N_{min} + 2$, that is, when the sample size in the middle between N_{min} and N_{max} only exceeds N_{min} by two (one per group), N_{mid} is returned as a result of the SSD algorithm. The arguments of this second function are identical to the ones of the first function (see Table 1).

Effect of number of subjects, frequency of observation, and study duration on power

In this section, we study the effect of the number of individuals (N), the frequency of observation (f), and the study duration (D) on power in the case of linear growth and equidistant time points. The subsequent simulations are executed with `m = 10000` datasets under each hypothesis in each iteration, an intercept variance of `var.u0 = 0.0333`, a slope variance of `var.u0 = 0.0333`, no covariance between the random effects (`cov = 0`), an error variance of `var.e = 0.02`, a standardized effect size of `eff.size = 0.8`, a fraction of information of `fraction = 1` resulting in $b = 1/N$ and a BF threshold

Table 1

Arguments to the function `BayeSSD`

Argument	Type	Meaning	Default value
`eta`	integer	η, the desired power level for both hypotheses	`0.8`
`BFthres`	integer	BF_{thres}, the threshold a BF must exceed in order to be considered convincing	`3`
`eff.size`	integer	δ, the expected standardized effect size	`0.8`
`m`	integer	number of datasets to be simulated in each iteration under each hypothesis	`1000`
`t.points`	vector	T_j, the time points of the measurement occasions, equal and unequal spacing possible	`c(0,1,2,3,4)`
`log`	logical	use log-linear growth?	`FALSE`
`var.u0`	integer	σ^2_{u0}, the intercept variance	`0.03`
`var.u1`	integer	σ^2_{u1}, the slope variance	`0.1`
`var.e`	integer	σ^2_e, the residual variance	`0.02`
`cov`	integer	σ_{u0u1}, the covariance between intercept and slope variance	`0`
`fraction`	integer	J, fraction of information to specify prior: $b = J/N$	1
`sensitivity`	logical	execute sensitivity analysis for different values (1, 2, 3) for `fraction`?	FALSE
`hyp`	string	perform SSD for which hypothesis, "h0"/"H0", "h1"/"H1" or "both"/"b"	`"both"/"b"`
`test`	string	which hypothesis should be compared against? If "alt", then $\mathcal{H}_0$ vs. $\mathcal{H}_1$ or vice versa, if "Hc"/"hc", then against the complement hypothesis ($\mathcal{H}_c$), if "Hu"/"hu", then against the unconstrained hypothesis ($\mathcal{H}_u$)	`"alt"`
`Neff`	string	if set to `"best"` then $N_{eff} = N * n$, if set to `"worst"` then $N_{eff} = N$	`"worst"`
`seed`	integer	set a seed for reproducibility	`NULL`

of `BFthres = 3`. Thus, in the case where $\mathcal{H}_1 : \beta_2 > 0$ is the true data-generating mechanism, $\beta_2 = \delta\sqrt{\sigma_{u1}} = 0.8 * \sqrt{0.001} = 0.0253$ (Raudenbush & Liu, 2001). On the other hand, if $\mathcal{H}_0 : \beta_2 = 0$ is the true data-generating mechanism, $\delta = 0$, obviously. Table 2 shows the effects of increasing D and f on the power of $\mathcal{H}_1$, while holding the number of subjects constant at $N = 100$. Note that Table 2 does not include power for $\mathcal{H}_0$ as this is largely unaffected by D and f. Table 3 illustrates the effects of increasing D and N for both hypotheses while keeping the frequency of observation constant at $f = 1$. Note that instead of an extensive simulation study resulting in large tables, we opt for a more contained simulation which showcases the most important operating characteristics of the

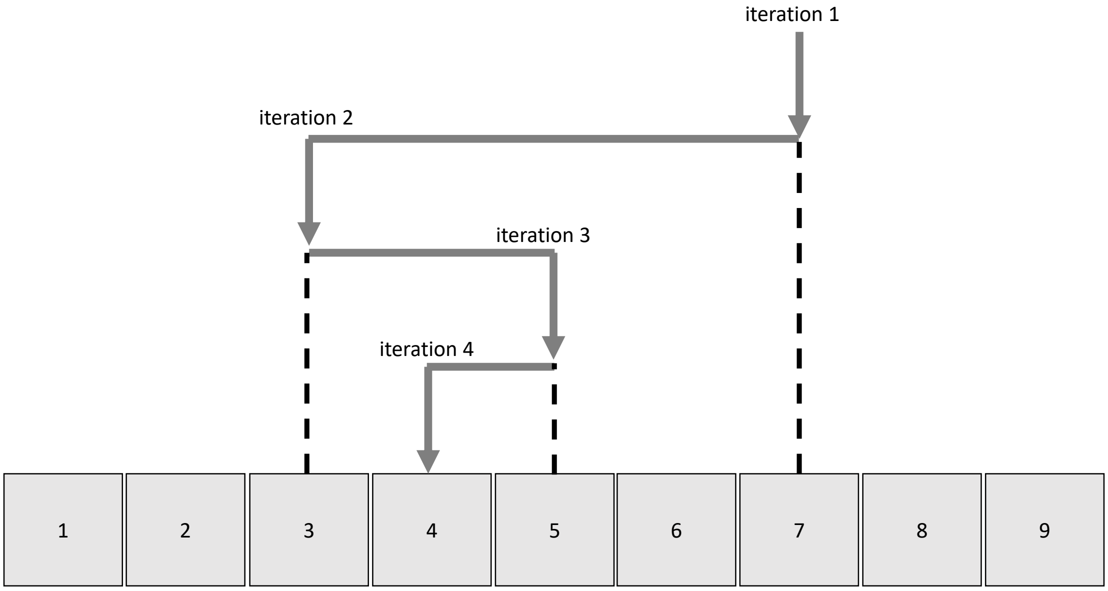


Figure 4
Procedure of a binary search algorithm.

power statistic in (22) in a clear and straightforward manner. In other words, in Table 2 and 3 we show that the Bayesian power behaves according to our expectations based on previous research on frequentist power in multilevel models (e.g., Moerbeek, 2008). For the same tables in case $\mathcal{H}_1$ is compared against $\mathcal{H}_c$ or $\mathcal{H}_u$, see the Appendix.

As can be seen in Tables 2 and 3, increasing one of the quantities f, D, and N consistently leads to higher power for $\mathcal{H}_1$. It can further be noted that increasing N is the most efficient strategy to improve power because the power level increases without bound while the effect of increasing f or D levels off at some point. This pattern is in line with previous findings on frequentist power in multilevel models (Moerbeek, 2008). However, while the power for $\mathcal{H}_0$ depends on N, it is largely unaffected by f and D.

Empirical Examples

In this section, two examples are introduced which are used to illustrate the SSD procedure. While both of these examples consist of "real" empirical longitudinal studies, we do not use their original datasets. Rather, we execute the SSD for a potential replication study with a certain desired power level. For that purpose, we use the author’s findings about estimates of fixed and random effects as "ingredients" for our SSD

Table 2

Effects of study duration (D) and frequency of observation (f) on Bayesian power

$\mathcal{H}_1 : \beta_2 > 0$			f			
D	1	2	3	4	5	6
1	-	.0438	.0479	.0490	.0507	.0548
2	.1351	.1572	.1872	.2189	.2496	.2696
3	.3003	.3743	.4432	.5019	.5442	.5860
4	.4930	.5870	.6655	.7077	.7551	.7787
5	.6557	.7382	.7927	.8258	.8494	.8664
6	.7531	.8228	.8533	.8759	.8895	.9015
7	.8205	.8733	.8963	.9093	.9155	.9198
8	.8579	.9016	.9146	.9208	.9277	.9320

Note. Number of individuals held constant at $N = 100$; number of measurement occasions $n = fD + 1$. $\mathcal{H}_1$ is compared against $\mathcal{H}_0 : \beta_2 = 0$.

Table 3

Effects of study duration (D) and number of individuals (N) on Bayesian power for $\mathcal{H}_0$ and $\mathcal{H}_1$

$\mathcal{H}_0$					N					
D	20	40	60	80	100	120	140	160	180	200
2	.6659	.7777	.8230	.8534	.8741	.8911	.8950	.9040	.9113	.9158
3	.6626	.7784	.8251	.8505	.8708	.8819	.8901	.8992	.9111	.9153
4	.6662	.7687	.8256	.8449	.8665	.8832	.8914	.8981	.9077	.9099
5	.6609	.7774	.8163	.8570	.8703	.8875	.8947	.9014	.9112	.9106
6	.6617	.7727	.8249	.8508	.8686	.8864	.8922	.9047	.9085	.9099
7	.6663	.7782	.8143	.8561	.8718	.8856	.8923	.9024	.9088	.9170
8	.6665	.7666	.8201	.8479	.8722	.8938	.8936	.8993	.9066	.9153
$\mathcal{H}_1$					N					
D	20	40	60	80	100	120	140	160	180	200
2	.0956	.0947	.1077	.1163	.1316	.1426	.1614	.1899	.2029	.2260
3	.1319	.1678	.2113	.2436	.2947	.3433	.3848	.4356	.4730	.5150
4	.1945	.2640	.3402	.4205	.4902	.5521	.6244	.6846	.7291	.7799
5	.2552	.3512	.4574	.5630	.6457	.7223	.7845	.8442	.8709	.9094
6	.3105	.4286	.5524	.6677	.7584	.8271	.8712	.9138	.9439	.9629
7	.3457	.4892	.6229	.7294	.8291	.8818	.9229	.9553	.9699	.9856
8	.3686	.5359	.6694	.7851	.8608	.9124	.9492	.9710	.9842	.9912

Note. Frequency of observations held constant at $f = 1$; $\mathcal{H}_0 : \beta_2 = 0$, $\mathcal{H}_1 : \beta_2 > 0$, number of measurement occasions $n = fD + 1$

algorithm. The two examples represent two typical cases: a) linear growth and equidistant time points and b) log-linear growth and non-equidistant time points.

Example 1: Antisocial thinking during adolescence: linear growth with equally spaced measurement occasions including sensitivity analysis

Similar to Raudenbush and Liu (2001) we use the study design of the National Youth Survey (NYS; Elliott, Huizinga, & Menard, 2012) to illustrate our method. Suppose a researcher aims to replicate the findings of the NYS using the BF and wants to determine the sample size corresponding to a user-selected power level. To determine that sample size, we use the findings of the original study about parameters such as the variances of the random effects as a starting point for our simulation. In the first cohort of the the NYS, adolescents in the US were interviewed yearly between 1976 and 1980 about their tolerance of antisocial behaviour. Because previous findings suggest a linear effect of time on the outcome in this age (Raudenbush & Chan, 1993), we use linear growth for the SSD. Furthermore, the distance between measurement occasions is constant, resulting in an equally spaced time vector. For a detailed description of the data, the interested reader is referred to Miyazaki et al. (2000).

Because in the original study there were no treatment and control condition, we add those to the design in order to make our method applicable to this example, analogous to Raudenbush and Liu (2001). We suppose that in the treatment condition, adolescents are educated about the consequences of antisocial behaviour while this is not the case in the control condition. Our aim is to test whether the difference in the rate of decline of the outcome (tolerance of antisocial behaviour) differs significantly between the treatment and the control condition. The corresponding hypotheses are the same as hypothesis $\mathcal{H}_0$ (6) and $\mathcal{H}_1$ (7) in the previous section.

The $n = 5$ measurement occasions are equally spaced resulting in a time vector of $T_j' = (0, 1, 2, 3, 4)$ in years. Decline over time is assumed to be linear, so the argument `log` is set to `FALSE` when running the function. Using the estimates from Raudenbush et al. (2001), we can fill in the expected variance of the residuals, intercept, and slope: $\sigma_e^2 = 0.0262$, $\sigma_{u0}^2 = 0.0333$, $\sigma_{u1}^2 = 0.0030$. The expected standardized effect size is

$\delta = 0.40$ corresponding to an expected coefficient of interaction $\beta_2 = \delta\sqrt{\sigma^2_{u1}} = 0.0219$. As we did not find the value for the covariance between the random effects, we assume that it is equal to zero. We decide to test our hypothesis $\mathcal{H}_1$ against the null hypothesis $\mathcal{H}_0$, so we set the `test` argument to `"alt"`. Next, we select a threshold value $BF_{thres} = 3$ and the desired power $\eta = .80$. Also, we ask for a sensitivity analysis for different values for the b fraction. Note that because the SSD procedure is a stochastic process, we need to set a seed for reproducibility. Now, we have all the ingredients for the SSD and feed them to the function as follows:

```
BayeSSD(m=10000, t.points=c(0,1,2,3,4), var.u0=0.0333,
var.u1=0.003, cov=0, var.e=0.0262, eff.size=.4, BFthres=3,
eta=.80, log.grow=FALSE, sensitivity=TRUE, seed=123)
```

The following output informs the user about the results of the SSD procedure.

```
The recommended sample size to achieve a power of at least 0.8
using b = 1 / N is N = 520
Power for H0: P(BF01 > 3 |H0) = 0.9507
Power for H1: P(BF10 > 3 |H1) = 0.8015


The recommended sample size to achieve a power of at least 0.8
using b = 2 / N is N = 477
Power for H0: P(BF01 > 3 |H0) = 0.9194
Power for H1: P(BF10 > 3 |H1) = 0.8084


The recommended sample size to achieve a power of at least 0.8
using b = 3 / N is N = 477
Power for H0: P(BF01 > 3 | H0) = 0.8999
Power for H1: P(BF10 > 3 | H1) = 0.8052
```

The output of the function informs us that we need to sample at least $N_1 = 520$ participants to achieve an at least 80% chance of making the correct inferential decision when using $b_{min} = \frac{1}{N}$ and $N_{2,3} = 477$ participants when using $b_2 = \frac{2}{N}$ or $b_3 = \frac{3}{N}$. Furthermore, we learn that the choice of the b fraction influences the resulting power notably. When using b_{min}, we need 43 more participants to achieve the desired power level compared to b_2 or b_3, whereas there is no difference between choosing b_2 or b_3 in this case. As mentioned before, the choice of b depends on whether a minimally informed prior (b_{min}) or a robust prior (b_2, b_3) is preferred. For a more extensive overview and guidelines for choosing b, see Gu et al. (2018).

These sample sizes might be considered high by the researcher and it is possible that available resources are insufficient for that large of a sample. As can be seen in Tables 2 and 3, there are options to obtain more power other than increasing N. One might consider increasing the study duration D or the frequency of observation f in order to achieve similar power with a smaller N.

Example 2: Systematic patient feedback in psychotherapy: log-linear growth and unequally spaced measurement occasions

For our second example, we use the study by Bovendeerd et al. (2022) as a starting point. In this study patients were allocated to either a treatment condition (therapy with feedback) or a control condition (therapy without feedback) and assessed at four occasions (week 0, 5, 13, and 638) with the Outcome Questionnaire (OQ-45; Lambert, Gregersen, & Burlingame, 2004). While it was the intention of the authors to measure every participant at the same occasion, this was not the case in practice. For the sake of illustrating our method for SSD, however, we assume that all subjects will be measured at the same points in time in the replication study. The data have a three-level structure (measurement, patient, therapist). Note that the amount of variance at the therapist level was very small as compared to the patient and measurement level. In order to illustrate our method in a simple, straightforward manner, we omit the third level (therapist) so that observations are nested only within patients. Again, we suppose that we perform SSD for a replication study and use the findings of Bovendeerd et al. (2022) as a starting point. The spacing between the measurement occasions is *unequal* and time is assumed to have a *log-linear* effect on the outcome.

In line with the results found by the authors, we expect that in the treatment condition, patients improve faster compared to the control condition (therapy without feedback). Note that a *lower* score on the outcome (OQ-45) corresponds with *more* well-being and *less* psychiatric symptoms (Bovendeerd et al., 2022). Therefore, we expect the growth to be smaller in the treatment group compared to the control group. Suppose that in this case, we want to compare our hypothesis to its complement $\mathcal{H}_c$, resulting in the following hypotheses:

$$\mathcal{H}_1 : \beta_2 < 0 \tag{24}$$

$$\mathcal{H}_c : \beta_2 \geq 0 \tag{25}$$

In this case, the $n = 4$ measurement occasions are irregularly spaced in time. Individuals were measured before, during, and after treatment. Additionally, growth over time is assumed to be log-linear, meaning that the resulting time vector is $T_j' = log(0, 5, 13, 638)$. We therefore set the `log` argument to `TRUE` and the `t.points` argument to `c(0, 5, 13, 638)`. Previous findings of Bovendeerd et al. (2022) provide us with the necessary estimates for our function: $\sigma_e^2 = 172,191$, $\sigma_{u0}^2 = 237.114$, $\sigma_{u1}^2 = 13.234$, $\sigma_{u0u1} = 18.282$. The authors' estimate of the interaction coefficient is $\beta_2 = -2.051$ corresponding to a standardized effect size of -0.564. Note that because $\mathcal{H}_1 : \beta_2 < 0$ instead of $\mathcal{H}_1 : \beta_2 > 0$, we have to reverse the sign (direction) of the effect size such that $\delta = 0.564$. Because we decide to test our hypothesis $\mathcal{H}_1$ against its complement hypothesis $\mathcal{H}_c$, we set the `test` argument to `"Hc"`. Also, because we are not interested in the power for $\mathcal{H}_0$, we can set the argument `hyp` to `"H1"`, reducing the computation time. Next, we establish a threshold value $BF_{thres} = 10$ and the desired power $\eta = .70$. Now, we have all the ingredients for the SSD and feed them to the function as follows:

```
BayeSSD(m=10000, t.points=c(0,5,13,638), var.u0=237.114,
var.u1=13.234, cov=18.282, var.e=172.191, eff.size=.564,
BFthres=10, eta=.70, log=TRUE, sensitivity=FALSE, seed=123,
test="Hc", hyp="H1")
```

The following output informs the user about the results of the SSD procedure.

```
The recommended sample size to achieve a power of at least 0.7
using b = 1 / N is N = 69
Power for H1: P(BF1c > 10 | H1) = 0.7043
```

This means that if $\mathcal{H}_1$ is indeed true, we need to sample at least $N = 69$ participants to achieve a probability of at least 70% obtain a BF_{1c} of at least 10 when using $b = \frac{1}{N}$.

Conclusion

In this paper, we present a method to perform SSD for Bayesian multilevel analysis with linear or log-linear growth via Monte-Carlo simulation. This method is implemented in the open-source R function `BayeSSD`. With this function researchers can perform Bayesian SSD tailored to their specific longitudinal trial and use the results to motivate their sample size in proposals for funding agencies and ethics committees. To limit the intense computational cost of running the simulation within the function, we developed a binary search algorithm that requires significantly less iterations, similar to Fu (2021). We demonstrate the use of this function based on two empirical examples of psychological longitudinal intervention studies. Furthermore, we show that the operating characteristics of the algorithm are clear-cut and in line with general expectations. The simulation results presented in this paper highlight that the basic concepts of frequentist power in multilevel models also apply in Bayesian power in these models: Increasing the number of level 2 observations (N) raises the power without bound whereas increasing the number of level 1 observations (n) raises the power only up to a certain point. Increasing the duration (D) of a study also boosts power, a finding that aligns with the results by Raudenbush and Liu (2001). However, in practice it is important to examine the associated costs as well as consequences for attrition which is a common occurrence in psychological longitudinal intervention studies. Longer studies with more measurement occasions tend to result in higher drop-out rates, resulting in a loss of power (Moerbeek, 2008). Trade-offs like these should be taken into account in the design phase of a study and further studies on Bayesian SSD taking attrition into account are needed.

The novelty of this research lies in the methods employed: Multilevel models are a flexible and viable method to evaluate the effectiveness of treatments longitudinal intervention studies. They have notable advantages over alternative approaches such as repeated measures ANOVA which have been highlighted in the introduction. The BF with its relatively straightforward interpretation is a versatile tool for evaluating hypotheses without relying on the often misunderstood *p*-values. Additionally, it provides one with the possibility of obtaining evidence *in favor of* the null hypothesis, something that is not possible with frequentist methods of NHST. The specific way of calculating the Bayes Factor via the AAFBF (Gu et al., 2018) does not require the researcher to specify a prior distribution, thereby removing some subjectivity from the inferential procedure and making the method more accessible for researchers who are less experienced in Bayesian statistics. Furthermore, calculating the AAFBF is much less computationally intensive compared to other BFs which employ complicated MCMC sampling algorithms to sample from the posterior. The Bayesian approach to statistical power in general enables us to move away from the unilateral approach of NHST methods and provides us with a new perspective on the concept of power.

However, the method presented in this study is not without limitations. The R function is not yet able to handle more than two informative hypotheses or more than one regression parameter. In the subsequent project, these limitations will be addressed among with the inclusion of a survival function to take different patterns of dropout into account when calculating power.

To our knowledge, this is the first study introducing an open-source software for Bayesian SSD in longitudinal trials. We hope that these results along with the R function `BayeSSD` accessible on GitHub enable researchers who use Bayesian multilevel modelling to carry out SSD in a straightforward manner.

Declarations

Funding

This work was funded by the NWO Open Competition SSH personal grant with a focus on scientific impact awarded to Dr. Mirjam Moerbeek. Grant number 406.21.GO.006.

Conflicts of interest

The authors have no competing interests to declare.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

Data sharing is not applicable to this article as no empirical data was collected or used.

Code availablilty

The entire R code containing the functions necessary to carry out the sample size determination algorithm as well as the code used for the creation of figures and tables can be found on the personal GitHub repository of the first author: https://github.com/ulrichlosener/BayesianSSD.

Authors' contributions

Ulrich Lösener: Development and implementation of the open-access software, generation and analysis of simulated data, creation of figures and tables, drafting and writing out the manuscript

Mirjam Moerbeek: Obtaining the NWO grant, conception and design of the simulation study, providing domain-specific expertise and guiding the study's focus, critical revision of the manuscript and software

Herbert Hoijtink: Providing relevant expertise and theoretical background, final approval of the manuscript

References

Althammer, S. E., Reis, D., Van der Beek, S., Beck, L., & Michel, A. (2021). A mindfulness intervention promoting work–life balance: How segmentation preference affects changes in detachment, well-being, and work–life balance. *Journal of Occupational and Organizational Psychology*, *94*(2), 282–308.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.

Bernardo, J. M., & Smith, A. F. (2009). *Bayesian theory* (Vol. 405). John Wiley & Sons.

Bosker, R., & Snijders, T. A. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* Sage.

Bovendeerd, B., De Jong, K., De Groot, E., Moerbeek, M., & De Keijser, J. (2022). Enhancing the effect of psychotherapy through systematic client feedback in outpatient mental healthcare: A cluster randomized trial. *Psychotherapy Research*, *32*(6), 710–722.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147.

Case, L. D., & Ambrosius, W. T. (2007). Power and sample size. *Topics in Biostatistics*, 377–408.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences.* Academic Press.

Conigliani, C., & O'Hagan, A. (2000). Sensitivity of the fractional bayes factor to prior distributions. *Canadian Journal of Statistics*, *28*(2), 343–352.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204–223.

Elliott, D. S., Huizinga, D., & Menard, S. (2012). *Multiple problem youth: Delinquency, substance use, and mental health problems.* Springer Science & Business Media.

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes.

Dental Press Journal of Orthodontics, *19*, 27–29.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379.

Fu, Q. (2022). *Sample size determination for bayesian informative hypothesis testing* (Unpublished doctoral dissertation). Utrecht University.

Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the bayesian t test and welch's test using the approximate adjusted fractional bayes factor. *Behavior Research Methods*, *53*(1), 139–152.

Gibbons, R. D., Hedeker, D., & DuToit, S. (2010). Advances in analysis of longitudinal data. *Annual Review of Clinical Psychology*, *6*, 79–107.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*(3), 647–674.

Goldstein, H., & Browne, W. (2003). *Multilevel models.* Arnold Publishers.

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143.

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. J. (2021). bain: Bayes factors for informative hypotheses [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=bain` (R package version 0.2.8)

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 229–261.

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., … others (2022). A review of applications of the bayes factor in psychological research. *Psychological Methods.*

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* Wiley-Interscience.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists.* CRC Press.

Hoijtink, H. (2022). Prior sensitivity of null hypothesis bayesian testing. *Psychological Methods*, *27*(5), 804.

Hoijtink, H., Gu, X., & Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, *72*(2), 219–243.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses* (Vol. 361). Springer.

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the bayes factor. *Psychological Methods*, *24*(5), 539.

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why bayesian psychologists should change the way they use the bayes factor. *Multivariate Behavioral Research*, *51*(1), 2–10.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications*. Routledge.

Hubbard, R. (2011). The widespread misinterpretation of p-values as error probabilities. *Journal of Applied Statistics*, *38*(11), 2617–2626.

Jeffreys, H. (1935). Some tests of significance, treated by the theory of Probability. *Proceedings of the Cambridge Philosophy Society*, *31*, 203–222.

Kaplan, R. M., Chambers, D. A., & Glasgow, R. E. (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and Translational Science*, *7*(4), 342–346.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020). Using bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *The Annals of Family Medicine*, *2*(3), 204–208.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The outcome questionnaire-45. In M. E. Maruish (Ed.), *The use of psychological testing for* *treatment planning and outcomes assessment: Instruments for adults* (3rd ed., p. 191–234). Lawrence Erlbaum Associates Publishers.

Lantz, B. (2013). The large sample size fallacy. *Scandinavian Journal of Caring Sciences*, *27*(2), 487–492.

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *The American Statistician*, *53*(2), 119–122.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244.

Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods*, *5*(1), 44.

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. *Journal of Educational and Behavioral Statistics*, *33*(1), 41–61.

Moerbeek, M. (2023). Bayesian sequential designs in studies with multilevel data. *Behavior Research Methods*, 1–13.

Moerbeek, M., & Teerenstra, S. (2015). *Power analysis of trials with multilevel data.* CRC Press.

Moerbeek, M., van Breukelen, G. J., & Berger, M. P. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*(4), 341–350.

Mulder, J. (2014). Prior adjusted default bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference*, *140*(4), 887–906.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 99–118.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., & Chan, W.-S. (1993). Application of a hierarchical linear model to the study of adolescent deviance in an overlapping cohort design. *Journal of Consulting and Clinical Psychology*, *61*(6), 941.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387.

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2023). What is a bayes factor? *Psychological Methods*, *28*(3), 705.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

Sekulovski, N., & Hoijtink, H. (2023). A default bayes factor for testing null hypotheses about the fixed effects of linear two-level models. *Psychological Methods*.

Tendeiro, J. N., & Kiers, H. A. (2019). A review of issues about null hypothesis bayesian testing. *Psychological Methods*, *24*(6), 774.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: surprising insights from bayes's theorem. *Psychological Review*, *110*(3), 526.

Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology*, *136*(3), 261–270.

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*, 87–102.

Van Buuren, S. (2018). *Flexible imputation of missing data.* CRC press.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S.

(2017). A systematic review of bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from `https://www.stats.ox.ac.uk/pub/MASS4/` (ISBN 0-387-95457-0)

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). *Bayesian evaluation of informative hypotheses*. Springer.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics*, 88–96.

Wang, F., & Gelfand, A. E. (2002). A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 193–208.

Appendix: Tables with comparisons against $\mathcal{H}_c$ and $\mathcal{H}_u$

Table 4

$\boldsymbol{\mathcal{H}_1 : \beta_2 > 0}$ *vs.* $\boldsymbol{\mathcal{H}_c : \beta_2 < 0}$

	f						
D	0.5	1	2	3	4	5	6
1		-	.4826	.4920	.5062	.5138	.5237
2		.7012	.7403	.7744	.8054	.8306	.8414
3		.8599	.9024	.9334	.9488	.9570	.9676
4		.9481	.9699	.9796	.9854	.9892	.9937
5		.9761	.9880	.9926	.9956	.9968	.9980
6		.9891	.9950	.9970	.9977	.9973	.9981
7		.9950	.9965	.9983	.9987	.9984	.9989
8		.9967	.9986	.9988	.9992	.9994	.9994

Note. Number of individuals held constant at $N = 100$; number of measurement occasions $n = fD + 1$.

Table 5

$\boldsymbol{\mathcal{H}_1 : \beta_2 > 0}$ *vs.* $\boldsymbol{\mathcal{H}_u : \beta_2}$

				f			
D	0.5	1	2	3	4	5	6
1		-	.2571	.2627	.2718	.2819	.2994
2		.4712	.5145	.5612	.5976	.6367	.6566
3		.6896	.7621	.8189	.8549	.8723	.8955
4		.8422	.8997	.9237	.9444	.9546	.9635
5		.9187	.9524	.9670	.9765	.9814	.9845
6		.9559	.9749	.9827	.9868	.9883	.9893
7		.9755	.9858	.9885	.9925	.9914	.9925
8		.9804	.9917	.9928	.9928	.9943	.9943

Note. Number of individuals held constant at $N = 100$; number of measurement occasions $n = fD + 1$.