Bayesian power equivalence in latent growth curve models

Stefan, A.M.; von Oertzen, T.

DOI
10.1111/bmsp.12193

Publication date
2020

Document Version
Final published version

Published in
British Journal of Mathematical & Statistical Psychology

the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2019), 73, 180–193

www.wileyonlinelibrary.com

Bayesian power equivalence in latent growth curve models

Angelika M. Stefan[1]* and Timo von Oertzen[2]

[1]Department of Psychology, University of Amsterdam, The Netherlands
[2]Department of Psychology, University of the Bundeswehr, Germany

Longitudinal studies are the gold standard for research on time-dependent phenomena in the social sciences. However, they often entail high costs due to multiple measurement occasions and a long overall study duration. It is therefore useful to optimize these design factors while maintaining a high informativeness of the design. Von Oertzen and Brandmaier (2013, *Psychology and Aging*, 28, 414) applied power equivalence to show that Latent Growth Curve Models (LGCMs) with different design factors can have the same power for likelihood-ratio tests on the latent structure. In this paper, we show that the notion of power equivalence can be extended to Bayesian hypothesis tests of the latent structure constants. Specifically, we show that the results of a Bayes factor design analysis (BFDA; Schönbrodt & Wagenmakers (2018, *Psychonomic Bulletin and Review*, *25*, 128) of two power equivalent LGCMs are equivalent. This will be useful for researchers who aim to plan for compelling evidence instead of frequentist power and provides a contribution towards more efficient procedures for BFDA.

1. Introduction

Researchers design experiments to gain knowledge of the world. In a world of limited resources, it is ethical to conduct these experiments efficiently (Halpern *et al.*, 2002). Hunter and Hoff (1967) define research efficiency as 'the amount of useful information obtained per unit cost'. Often, longitudinal studies entail especially high costs. These accrue either due to a long overall study duration, for example when a treatment has to be administered over a long period of time, or due to a large number of measurement occasions, for example when non-reusable testing material is spent at each testing event. It is therefore especially important to plan longitudinal studies carefully so that an optimal balance between study costs and the expected gain in information can be achieved (Brandmaier *et al.*, 2015).

Longitudinal designs can be statistically evaluated with a sub-group of structural equation models (SEMs; for an overview see e.g., Baltes *et al.*, 1988) called Latent Growth Curve Models (LGCMs; see e.g., Duncan & Duncan, 2009). In a simple LGCM, the values of a variable across several measurement occasions (x_i) are modeled as a combination of a

*Correspondence should be addressed to Angelika M. Stefan, Department of Psychology, Faculty of Behavioral and Social Sciences, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018WS Amsterdam, The Netherlands (email: a.m.stefan@uva.nl).

DOI:10.1111/bmsp.12193

latent intercept (*I*) and a latent slope (*S*). The intercept has a constant influence on the measurement occasions, while the slope adds time-dependent linear changes (see Figure 1). To add nonlinear changes, a quadratic or higher-order term can be introduced (Duncan & Duncan,2009). For example, Lindenberger and Ghisletta (2009) investigated cognitive and sensory decline in elderly participants with an LGCM. In this context, the intercept parameter captured the participants' initial abilities and the slope parameter captured the extent of the linear time-dependent decline.

An advantage of LGCMs is that they allow the direct estimation of between-subjects variability in the latent intercept and slope, described as the variance of the intercept (σ_I^2) and the variance of the slope (σ_S^2) in the model. These random effects represent the individual differences in initial performance and change, respectively (Rogosa & Willett, 1985). In an LGCM where the intercept reflects the initial status of the observed variable, the intercept-slope covariance (σ_{IS}) reflects the extent to which individual differences in the initial status correlate with subsequent change (Rovine & Molenaar, 1999). Thus, in the example used earlier (Lindenberger & Ghisletta, 2009), the variance of the intercept can be interpreted as the variability of cognitive and sensory abilities of participants at the beginning of the study. The variance of the slope corresponds to differences in the steepness of the cognitive decline between participants. A positive covariance between intercept and slope in the example would show that participants with higher initial abilities suffer from a more rapid decline.

In a frequentist setting, an important aspect of the quality of a design is its statistical power, which is defined as the long-term probability of correctly rejecting the null hypothesis under a given population effect size that differs from zero. The statistical power of a design depends on the size of the effect in the population, on the significance level α of the hypothesis test, on the sample size *N*, and on the measurement design

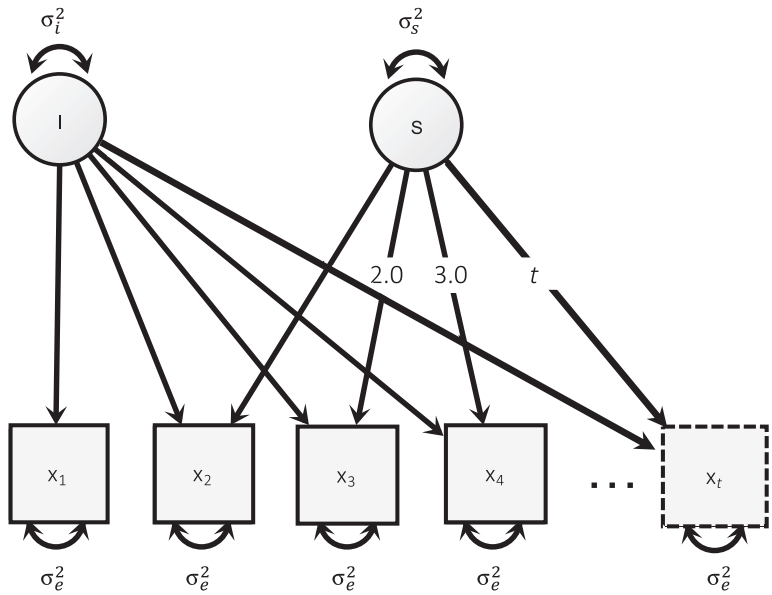

Figure 1. Schematic representation of a Latent Growth Curve Model (LGCM). More measurement occasions can be added as depicted for x_t. Latent variables represent the intercept (*I*) with variance σ_I^2 and slope (*S*) with variance σ_S^2. Figure available under a CC-BY4.0 license at https://osf.io/hkt4p/.

(Brandmaier, *et al.*, 2018; Cohen, 1992). For most traditional hypothesis tests, such as a *z*-test or a *t*-test, it is possible to calculate the statistical power analytically (Murphy *et al.*, 2014). However, for most SEMs there is no analytical solution available, so the statistical power of a model has to be estimated via numerical approximations (e.g., Saris & Satorra, 1993) or through simulations (e.g., Hertzog, *et al.*, 2008, Muthén & Muthén, 2002).

Von Oertzen (2010) introduced the concept of power equivalence, which describes that two designs have the same statistical power to detect a true effect. Power equivalence can be used to find research designs that are most resource efficient among designs with the same power. For example, von Oertzen and Brandmaier (2013) illustrated how power equivalence facilitates finding a cost-optimal solution among multiple longitudinal designs. In longitudinal designs, power equivalence can be established by balancing the overall duration of the study and the number of measurement occasions. To keep the power constant, more measurement occasions are required if the overall study duration is shortened. By comparing multiple power-equivalent longitudinal designs based on data and cost estimates from the Berlin Aging Study (BASE; Ghisletta, *et al.*, 2006), von Oertzen and Brandmaier (2013) showed that the overall study costs could be reduced by 16% compared to the original design while keeping the statistical power with respect to the variance of slopes constant. Thus, power equivalence can facilitate the planning of future studies in two ways: First, instead of conducting multiple potentially resource-intensive power analyses for different designs, a power analysis has to be computed only once for a theoretically infinite number of power-equivalent designs. Second, knowing that certain designs do not differ in an important aspect of design quality, researchers can focus on minimizing the costs (Hunter & Hoff, 1967).

Conceptually, power equivalence as applied in von Oertzen and Brandmaier (2013) can be described by the following procedure. Any LGCM can be reduced to a power equivalent model with a minimum number of observed parameters, from which further power equivalent models can be derived. These power equivalent models balance different design parameters, for example the number of measurement occasions ($j = 1, \ldots, k$) and the time distance between measurement occasions, modeled in the path parameters $\theta_S \rightarrow x_j$, so that the power to detect an effect (e.g., $\sigma^2_S > 0$) is equivalent.[1] This is reflected in the effective error variance σ^2_{eff} which is shared by all power-equivalent models. Figure 2 schematically depicts this trade-off: A linear trend is measured with three power equivalent designs which differ in their number of measurement occasions and their overall study duration.

In recent years, the replicability crisis (Pashler & Wagenmakers, 2012) as well as continuing criticism regarding the frequentist hypothesis testing framework (e.g., Edwards, *et al.*, 1963; Wagenmakers, 2007) have led to a growing interest in Bayesian methods for statistical inference.[2] The single most important quantity in Bayesian hypothesis testing is the Bayes factor (Kass & Raftery, 1995). Mathematically, the Bayes factor (BF_{10}) is defined as the ratio of the marginal likelihood of the data under the alternative model ($p(\mathbf{D}|H_1)$) and the marginal likelihood of the data under the null model ($p(\mathbf{D}|H_0)$). It provides a continuous quantification of the evidence in favor of one statistical model compared to another statistical model.

[1] Note that although j in theory can go down to $j = 1$, in practical cases j needs to be at least equal to the number of latent variables (e.g., two for a linear LGCM, or three for a quadratic LGCM) to estimate all distribution parameters of the latent variables.

[2] An easily accessible introduction to Bayesian inference can be found in Etz & Vanderkerckhove (2018).

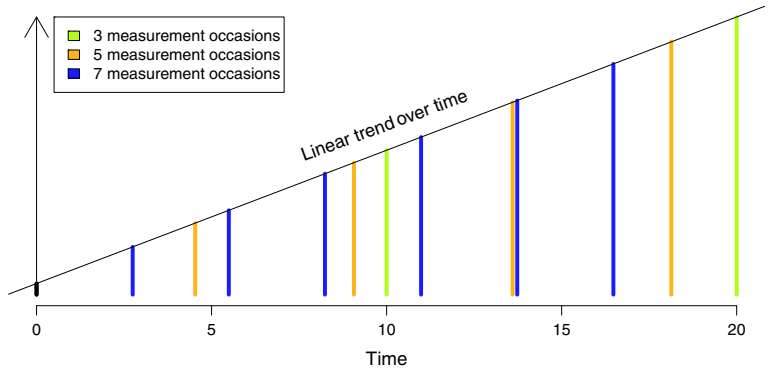

Figure 2. Measurement occasions of three power equivalent models measuring a linear trend. The power equivalent models were computed for $\sigma_i^2 = 2$, $\sigma_e^2 = 1$, and three, five, and seven measurement occasions. All designs assume a first measurement occasion at time $t = 0$. Figure available under a CC-BY4.0 license at https://osf.io/hkt4p/.

Since most researchers aim to collect compelling evidence in a study, both very large or very small Bayes factors can be regarded as a desirable outcome of a study. For example, a Bayes factor of $BF_{10} = 10$ indicates a tenfold increase in prior odds to posterior odds in favor of the alternative hypothesis after having observed the data, while a Bayes factor of $BF_{10} = 1/10$ indicates a tenfold increase in prior odds in favor of the null hypothesis. How large the Bayes factors get that an experiment yields, depends on the tested models (described by likelihoods and prior distributions), on the population effect size, on the amount of collected data, that is, the number of observations in the sample, and on the measurement design (Stefan, *et al.*, 2019). Assuming that the models are determined by the research question, only the sample size and measurement design can be directly influenced by the researcher. This shows that researchers who use Bayesian statistics to evaluate their data are also in the need to balance the costs and the information gain of their designs – in other words that design planning is an important topic from a Bayesian viewpoint, too.

How can researchers find an adequate sample size or measurement design so that their study likely yields compelling evidence, but is also designed economically? Schönbrodt & Wagenmakers (2018) proposed a framework called 'Bayes Factor Design Analysis' (BFDA) that enables researchers to find the expected Bayes factors of their design. Their approach is based on Monte Carlo simulations where data are repeatedly simulated under a population model ('design prior') and a Bayesian hypothesis test is conducted for each of these samples. BFDA is applicable to both sequential Bayesian designs, where the sample size is gradually increased until a prespecified Bayes factor is reached, and fixed-N designs, where the sample size is specified prior to data collection. For the latter more traditional sampling procedure, a BFDA results in a distribution of Bayes factors that enables researchers to assess the informativeness of their planned design.

In this paper, we show that the notion of power equivalence can be extended to Bayesian hypothesis tests. Specifically, we show that the results of a BFDA for a fixed-*N* design (Schönbrodt & Wagenmakers, 2018) of two power equivalent models as defined by von Oertzen (2010) are equivalent. Our findings are not only relevant on a conceptual level as they instantiate a bridge between frequentist and Bayesian methods. They also provide Bayesians with a possibility of design justification in longitudinal settings and help to save resources in design planning because computationally expensive BFDAs need to be conducted only once for power equivalent designs.

Our paper is structured as follows: First, we will formally prove the equivalence of BFDA results for power equivalent models. In a second step, we will substantiate our proof with a simulation for power equivalent LGCMs. Then, we will provide an application example that illustrates how Bayesian power equivalence can facilitate design planning. We will discuss the implications and limitations of our findings at the end of this article.

2. Formal proof of BFDA equivalence for power equivalent models

In this section, we show formally that two power equivalent models with the same parameter set θ will also produce the same distribution of the Bayes Factor when comparing two hypotheses about θ under data generated by a population model. We assume that both hypotheses are given by a prior distribution π_1 and π_1 for θ, where as usual one or both can be point hypotheses, i.e., degenerated prior distributions with the mass fixed at any specific point.

Power equivalence on multivariate normal models, as defined in von Oertzen (2010), can be expressed as a combination of two basic power equivalent operations. The first one is a linear transformation of the observed variables, the second an omission of observed variables with a probability distribution which is constant with respect to θ, and which are independent of other variables. For example, in an LGCM, the linear transformation transforms the measurement model into a minimal model with one observed variable that is dependent on the latent slope and a number of variables that are independent of the latent slope (and hence of the slope variance parameter). An example for a power equivalent transformation of an LGCM can be seen in Figure 3. The mathematical details of the calculation can be found in the Appendix.

Let (S,m) be the estimated covariance matrix and mean of a sample and (Σ, μ) of a model. In the following, we will write $\mathcal{L}_{\Sigma,\mu}(S, m)$ for the minus two log likelihood, i.e.,

$$\mathcal{L}_{\Sigma,\mu}(S, m) = -2 \log L(S, m|\Sigma, \mu).$$

We start by showing two simple lemmas.

Lemma 1. *For any multivariate normal model with covariance matrix Σ and mean μ, an orthogonal transformation Q on the model space does not change the likelihood function.*

Proof. The minus two log likelihood of a multivariate normal with parameter μ and Σ and a dataset with mean m and covariance matrix S per participant is

$$\mathcal{L}_{\Sigma,\mu}(S, m) = c + \ln(|\Sigma|) + \mathrm{Tr}(\Sigma^{-1}S) + (m - \mu)^T\Sigma^{-1}(m - \mu).$$

Transforming all four distribution parameters with Q results in

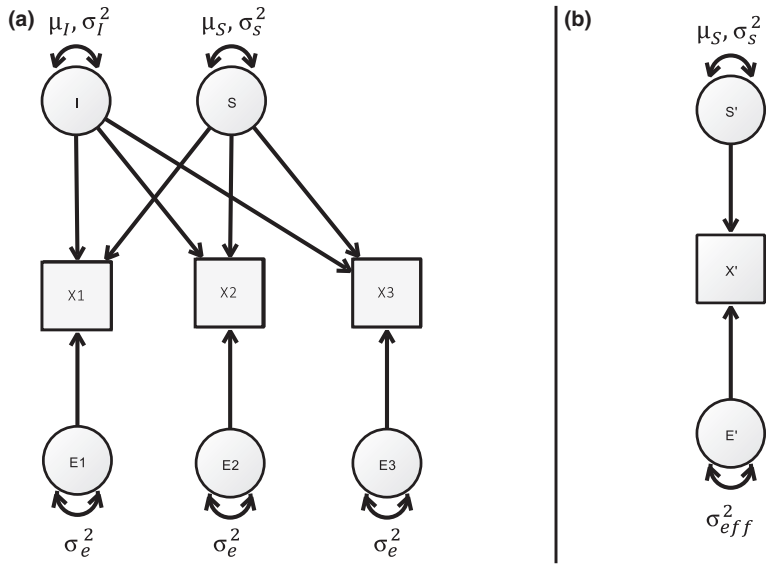

Figure 3. Power equivalent reduction of an LGCM. Panel A shows an LGCM with three measurement occasions which can be reduced to the minimal power equivalent model displayed in panel B. Figure available under CC-BY4.0 license at https://osf.io/hkt4p/.

$$\begin{aligned}\mathcal{L}_{Q\Sigma Q^T, Q\mu}(QSQ^T, Qm) &= c + \ln(|Q\Sigma Q^T|) + \mathrm{Tr}(Q\Sigma^{-1}Q^TQSQ^T) \\ &\quad +(m-\mu)^TQ^TQ\Sigma^{-1}Q^TQ(m-\mu) \\ &= c + \ln(|Q\Sigma Q^T|) + \mathrm{Tr}(Q\Sigma^{-1}SQ^T) + (m-\mu)^T\Sigma^{-1}(m-\mu),\end{aligned}$$

where the determinant and the trace do not change by an orthogonal transformation, therefore,

$$\begin{aligned}-2\log L(QSQ^T, Qm|Q\Sigma Q^T, Q\mu) &= c + \ln(|\Sigma|) + \mathrm{Tr}(\Sigma^{-1}S) + (m-\mu)^T\Sigma^{-1}(m-\mu) \\ &= \mathcal{L}_{\Sigma,\mu}(S, m).\end{aligned}$$

□

Lemma 2. *For any multivariate normal model with covariance matrix* Σ *and mean* μ*, omitting observed variables which have distributions that are constant with respect to some parameter set* θ *and are independent of all other parameters does not change the likelihood ratio of any two parameter values* θ_1 *and* θ_2*.*

Proof . For simplicity of notation, we prove that the difference of the minus two log likelihoods is constant. Let $\Sigma = \begin{pmatrix} \Sigma_1(\theta) & 0 \\ 0 & \Sigma_2 \end{pmatrix}$ be the separation of Σ and $\mu = \begin{pmatrix} \mu_1(\theta) \\ \mu_2 \end{pmatrix}$ be the separation of μ into a first part that depends on θ and a second, independent part that does not. We separate the data distribution accordingly. Note that the covariances between the two blocks in the data distribution are not relevant for the likelihood, i.e., we can write

$$\mathcal{L}_{\Sigma(\theta),\mu(\theta)}(S,m) = c + \ln(|\Sigma_1(\theta)|) + \mathrm{Tr}(\Sigma_1(\theta)^{-1}S_1) + (m_1 - \mu_1(\theta))^T\Sigma_1(\theta)^{-1}(m_1 - \mu_1(\theta)) \\ + \ln(|\Sigma_2|) + \mathrm{Tr}(\Sigma_2^{-1}S_2) + (m_2 - \mu_2)^T\Sigma_2^{-1}(m_2 - \mu_2)$$

When taking the difference of the minus two log likelihoods for θ_1 and θ_2, the second part of the equation and c cancels, so that the difference solves to

$$\begin{aligned}\mathcal{L}_{\Sigma(\theta_1),\mu(\theta_1)}(S,m) - L_{\Sigma(\theta_2),\mu(\theta_2)}(S,m) &= \ln(|\Sigma_1(\theta_1)|) + \mathrm{Tr}(\Sigma_1(\theta_1)^{-1}S_1) \\ &\quad +(m_1 - \mu_1(\theta_1))^T\Sigma_1(\theta_1)^{-1}(m_1 - \mu_1(\theta_1)) \\ &\quad -\ln(|\Sigma_1(\theta_2)|) - \mathrm{Tr}(\Sigma_1(\theta_2)^{-1}S_1) \\ &\quad -(m_1 - \mu_1(\theta_2))^T\Sigma_1(\theta_2)^{-1}(m_1 - \mu_1(\theta_2)) \\ &= \mathcal{L}_{\Sigma_1(\theta_1),\mu_1(\theta_1)}(S_1,m_1) - \mathcal{L}_{\Sigma_1(\theta_2),\mu_1(\theta_2)}(S_1,m_1).\end{aligned}$$

□

We conclude that the likelihood ratio remains constant under both base power equivalent operations, and hence under all combinations of those. Since the Bayes factor is the ratio of two prior-weighted likelihoods, we conclude further that the Bayes factor is unaltered by power equivalent transformations for any data set (S,m) and parameter sets θ_1 and θ_2. Thus, in particular, the distribution of the Bayes factor is identical for any priors π_1 and π_2 and any data distribution:

Corollary 3. *If* $(\Sigma_A(\theta), \mu_A(\theta))$ *and* $(\Sigma_B(\theta), \mu_B(\theta))$ *are two power equivalent multivariate normal models A and B, then under any distribution for data sets (S,m) and prior distributions* π_1 *and* π_2 *to be compared, the corresponding distribution of the Bayes factor is identical for both models.*

Proof. For simplicity, we omit the explicit separation of S and m in S_1 and S_2 and m_1 and m_2, respectively, because the irrelevant parts are ignored by the likelihood function (see proof of Lemma 2). For any specific outcome (S,m) of the random variable representing the data, let (S^*, m*) be the power equivalent transformation of the data as explained at the beginning of this section. The Bayes factor for the first model is given by

$$\mathrm{BF}_{A_{12}}(S,m) = \frac{\int_{\theta_1} L(S,m|\Sigma_A(\theta_1),\mu_A(\theta_1))\pi_1(\theta_1)\mathrm{d}\theta_1}{\int_{\theta_2} L(S,m|\Sigma_A(\theta_2),\mu_A(\theta_2))\pi_2(\theta_2)\mathrm{d}\theta_2}$$

which can be rewritten as

□

$$
\begin{aligned}
\mathrm{BF}_{A_{12}}(S, \boldsymbol{m}) &= \int_{\theta_1} \frac{\boldsymbol{L}(S, \boldsymbol{m}|\Sigma_A(\theta_1), \mu_A(\theta_1))\pi_1(\theta_1)}{\int_{\theta_2} \boldsymbol{L}(S, \boldsymbol{m}|\Sigma_A(\theta_2), \mu_A(\theta_2))\pi_2(\theta_2)\mathrm{d}\theta_2}\mathrm{d}\theta_1 \\
&= \int_{\theta_1} \frac{1}{\int_{\theta_2} \frac{\boldsymbol{L}(S, \boldsymbol{m}|\Sigma_A(\theta_2), \mu_A(\theta_2))\pi_2(\theta_2)}{\boldsymbol{L}(S, \boldsymbol{m}|\Sigma_A(\theta_1), \mu_A(\theta_1))\pi_1(\theta_1)}\mathrm{d}\theta_2}\mathrm{d}\theta_1 \\
&= \int_{\theta_1} \frac{1}{\int_{\theta_2} \frac{\boldsymbol{L}(S^*, \boldsymbol{m}^*|\Sigma_B(\theta_2), \mu_B(\theta_2))\pi_2(\theta_2)}{\boldsymbol{L}(S^*, \boldsymbol{m}^*|\Sigma_B(\theta_1), \mu_B(\theta_1))\pi_1(\theta_1)}\mathrm{d}\theta_2}\mathrm{d}\theta_1 \\
&= \frac{\int_{\theta_1} \boldsymbol{L}(S^*, \boldsymbol{m}^*|\Sigma_B(\theta_1), \mu_B(\theta_1))\pi_1(\theta_1)\mathrm{d}\theta_1}{\int_{\theta_2} \boldsymbol{L}(S^*, \boldsymbol{m}^*|\Sigma_B(\theta_2), \mu_B(\theta_2))\pi_2(\theta_2)\mathrm{d}\theta_2} \\
&= \mathrm{BF}_{B_{12}}(S^*, \boldsymbol{m}^*).
\end{aligned}
$$

Since the Bayes factor is identical for both models for any specific outcome of the data, its distribution under any random distribution of (S,m) is identical for both power equivalent models.

3. Simulation study

We performed a simulation study to illustrate the equivalence of Bayes factor distributions for power equivalent LGCMs. As in von Oertzen and Brandmaier (2013), we concentrated on a single parameter of interest: σ_S^2, the interindividual variance in the latent slope parameter. The focal Bayesian hypothesis test therefore compared the two hypotheses $\mathcal{H}_0$: $\sigma_S^2 = 0$ and $\mathcal{H}_1 : \sigma_S^2 \sim \pi_1$ where π_1, is a prior distribution that allows the parameter σ_S^2 to vary. We operationalized this prior distribution as a gamma distribution with a shape parameter of $k = 1$ and a rate parameter of $\beta = 0.5$. This prior places most weight on parameter values between 0 and 6 and can be considered as an example for an informed prior for typical effect sizes in psychology (see e.g., Duncan, *et al.*, 2006; Iddekinge, *et al.*, 2009; von Oertzen & Brandmaier, 2013). In this special case, all parameters of the model apart from σ_S^2 are considered to be known and fixed. Thus, the Bayes factor can be calculated through a simple integration procedure.

For our simulation study, we conducted a total of 36 BFDAs, where each BFDA result is based on 1000 Bayes factors. All BFDAs were performed using the following Monte Carlo simulation algorithm:

1. Find three power equivalent models with the given parameters for σ_E^2 and σ_I^2;
2. simulate 1,000 datasets for each of the models given a certain population parameter (design prior) for σ_S^2;
3. compute the Bayes factor for each of the datasets.

We compare the results of a fixed-N BFDA for 3 power equivalent LGCMs under 12 different population models (design priors). The three power equivalent models have 7, 5, and 3 equally distanced measurement occasions, respectively, and were computed using the equations provided in von Oertzen and Brandmaier (2013; see the Appendix below). In the simulations, we varied the variance of the intercept σ_I^2, the residual variance σ_E^2, and the true variance of the slope ($\sigma_S^2 \mid H1$). All BFDAs were conducted for a sample size of $N = 300$.

Figure 4 shows the distributions of log Bayes factors for the three power equivalent models under all simulated conditions. Overall, the distributions are nearly identical for

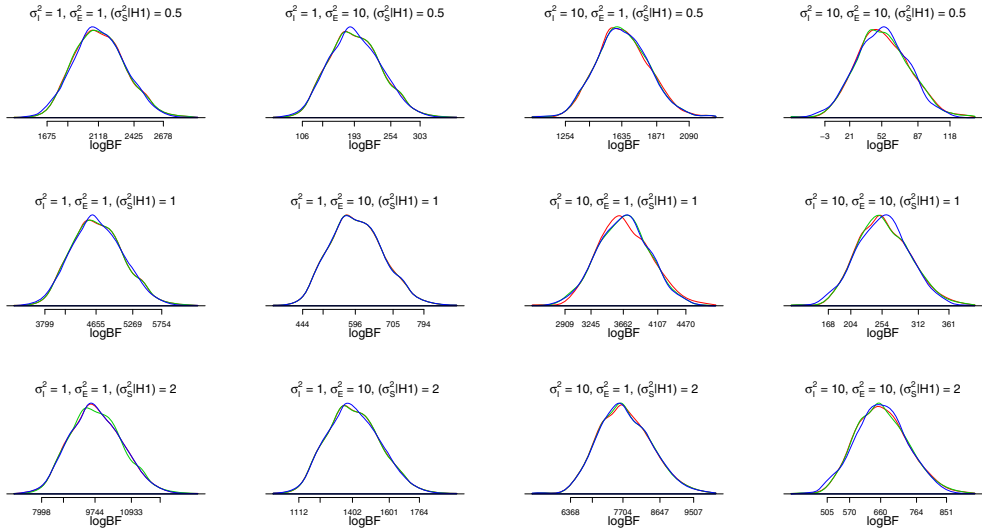

Figure 4. BFDAs for power equivalent models yield almost identical results. Figure shows the distribution of the log Bayes factors for power equivalent models with 3 (colored red), 5 (colored green), and 7 (colored blue) measurement occasions. Simulations were conducted with 1,000 iterations and different population parameters for the variance of the intercept σ_i^2, error variance σ_e^2, and variance of the slope σ_s^2 on a fixed sample size of $N = 300$. Figure available under a CC-BY4.0 license at https://osf.io/hkt4p/.

the power equivalent models which illustrates the formal proof of BFDA equivalence conducted in the previous section of this article. Generally, the Bayes factors are very large, which happens due to the relatively large dataset and the assumption that several important parameter values of the model are already known. There are small differences in the Bayes factor distributions that can be explained through the random variation in the simulation process.

The simulation code as well as the simulation results are openly accessible on https://osf.io/hkt4p/.

4. Application example: Effects of a mindfulness training

In this section, an applied example is discussed that illustrates how the notion of power equivalence can be used to facilitate *a-priori* design analyses for longitudinal studies. We will build on a study by Kiken *et al* (2015) who investigated the psychological effects of a mindfulness training. Mindfulness is a cognitive state of nonjudgmental awareness in which an individual pays attention to the thoughts, emotions, and sensations of the moment. Kiken *et al* (2015) measured state mindfulness with the Toronto Mindfulness Scale (Lau *et al*.,2006) at seven equally distanced measurement occasions during an ongoing mindfulness training that was directed at increasing the participants' general level of mindfulness. Using an LGCM, they concluded that while the training led on average to an increase in mindfulness, there were noticeable differences between individuals regarding the amount of change, i.e., there was considerable variability in the slope of state mindfulness.

Table 1. Power-equivalent models for testing the variance of slopes in a mindfulness training based on results of Kiken *et al.* (2015)

Waves	Assessment time	Wave costs	Running costs	Total costs
3	7.28	1,500	3,642	5,142
7	6.00	3,500	3,000	6,500
10	3.55	5,000	1,773	6,773

In this example application, we assume that researchers developed a new training method that is supposed to be equally effective for all participants. As the researchers would like to quantify evidence in favor of the null hypothesis ($\sigma_S^2 = 0$), they decide to use Bayesian hypothesis testing (Wagenmakers *et al*, 2018). When planning the study, they have two goals: Making sure that their envisioned sample size is large enough to obtain strong evidence in favor of the null hypothesis ($BF_{01} \geq 10$) if the null hypothesis is true and minimizing the overall study costs. For this example, we roughly estimate that the costs for each measurement occasion are \$10 per participant (e.g., for participant compensation or data entry), and that the running costs are \$500 per week (e.g., for renting lab space, employing assistants to run the study). We further assume that the envisioned sample size of the researchers is $N = 50$. Thus, when planning the study, two design questions come up: (1) Is a sample size of $N = 50$ enough to achieve strong evidence in favor of the null

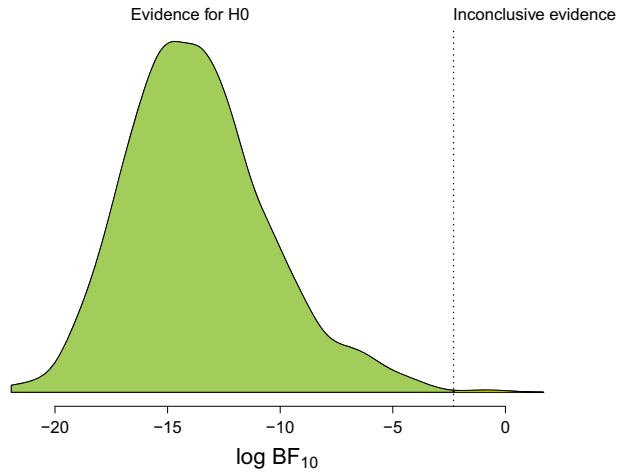

Figure 5. Bayes factor distribution resulting from a BFDA based on the results of Kiken et al. (2015) for a design with a fixed sample size of $N = 50$ and a true population effect size of $\sigma_S^2 = 0$. Figure available under a CC-BY4.0 license at https://osf.io/hkt4p/.

hypothesis when the null hypothesis is true, and (2) which of the power equivalent designs is most cost-efficient?

First, the researchers can now conduct a BFDA based on the design and results of the original study, that is seven equally distanced measurement occasions, a variance of intercepts of $\sigma_I^2 = 43.6$, and an error variance of $\sigma_E^2 = 21.45$. The results for a sample size of $N = 50$ show that the Bayes factor (BF_{10}) will be smaller than 0.1 in 99.8% of the cases,

that is, there is a high chance to obtain strong evidence in favor of the null hypothesis if the null hypothesis is true (see Figure 5). Being convinced by the high degree of informativeness, the researchers can now proceed to find the most cost-efficient design with the same power. Using power equivalence, the researchers can come up with several power-equivalent measurement designs. Table 1 shows three power-equivalent designs with 3, 7, and 10 measurement occasions, respectively (see the Appendix for details about the computation). All these designs share the same Bayes factor distribution based on the BFDA of the original design. However, they differ in their respective costs. As we can see from the total costs in Table 1, the measurement design with three measurement occasions is the most cost-efficient. Prolonging the overall study duration by 1.2 weeks, but reducing the number of measurement occasions to three can therefore lead to a cost reduction of roughly 20%. There is no need to recalculate the BFDA because the researchers already know that all power-equivalent designs are equally informative.

5. Discussion

Reducing study costs while keeping the results informative is an important practical objective of experimental design (Hunter & Hoff, 1967). In longitudinal studies, a cost reduction can often be achieved by finding a trade-off between the total duration of the study and the number of measurement occasions. In a frequentist setting, researchers can optimize this trade-off while keeping the design informative by comparing several power equivalent models (von Oertzen, 2010; von Oertzen & Brandmaier, 2013). While these models all have the same statistical power (Cohen, 1992), they exhibit different combinations of overall study length and number of measurement occasions. In this paper, we showed that the notion of power equivalence can be transferred to a Bayesian hypothesis testing framework. Specifically, we could show that power equivalence models yield the same Bayes factor distributions in a Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers 2018). Therefore, power equivalent designs are equally informative both from a frequentist and Bayesian viewpoint. This shows that power equivalent models can also be used in Bayesian design planning to negotiate trade-offs between costs and informativeness in longitudinal studies.

Our findings can be interpreted as an extension of both power equivalence (von Oertzen & Brandmaier, 2013; von Oertzen, 2010) and BFDA (Schönbrodt & Wagenmakers, 2018). From the perspective of power equivalence, we provide a straightforward generalization of the approach and show that it can also be used in the Bayesian process design planning. This highlights the relevance of the approach and raises the question whether further generalizations are possible. For example, the general notion of power equivalence could be generalized to statistical models other than Latent Growth Curve Models (LGCMs). Our results show that this would be a relevant contribution to design planning methods both from a frequentist and a Bayesian viewpoint. From the perspective of BFDA, our findings provide a first step towards a simplification of the procedure. Since the approach is based on Monte Carlo simulations, conducting a BFDA can be computationally expensive. Finding models that yield the same BFDA results can substantially facilitate the process of Bayesian design planning because a BFDA needs to be conducted only once for all of these power equivalent models. Our results show that finding such power equivalent models is possible. Future research could be directed at finding more conditions for equality of BFDA results and to extend our results to sequential Bayesian designs.

By making power equivalence available to a new statistical domain, our study increases its practical applicability to the planning of experimental designs. Additionally, we make it easy for researchers to optimize their study designs based on power equivalence and BFDA by providing the code for all analyses conducted in this paper online (see https://osf.io/hkt4p/). By using well-documented functions, we hope to encourage researchers to reuse our code and adapt it to their own practical applications. However, currently, the practical applicability of power equivalence in experimental design is still restricted by two important limitations. Firstly, the mathematical derivation for power equivalence requires that parameters which are not part of the hypothesis (in our example the variance of the intercept σ_I^2 and the error variance σ_E^2) are fixed. In practice, this is a strong assumption. However, if these parameters are not known, they (or the effective error σ_{eff}^2) can be estimated prior to the computation of power equivalence. A second limitation is that currently power equivalence requires a fixed structure matrix (von Oertzen, 2010), so it is only directly applicable to models like LGCMs, Change Score Models, Dual Change Score Models, Latent Differential Models, and basic models (e.g., ANOVAs). Nevertheless, these describe a considerable part of SEMs used today.

From a broader perspective, our findings illustrate that despite of methodological differences and occasional heated debates between frequentist and Bayesian methods and their respective proponents (see e.g., Wagenmakers, *et al.*, 2008), often relevant insights can be gained from describing the world from both perspectives. We hope that by showing how the notion of power equivalence and the BFDA method can be combined, we will have made a contribution towards an increased feasibility of Bayesian experimental planning. Eventually, we hope that the existence of straightforward methods for design planning can encourage more researchers to plan their study designs for efficiency and informativeness.

Acknowledgements

We want to thank Andreas Brandmaier and Eric-Jan Wagenmakers for their comments on an earlier draft of this paper. This research was supported by an NWO grant to AS (406.18.556).

References

Baltes, P., Reese, H., & Nesselroade, J. (1988). *Life-span developmental psychology: Introduction to research methods* (reprint of the 1977 edition). Hillsdale, NJ: Erlbaum.

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Hertzog, C., & Lindenberger, U. (2015). Lifespan: A tool for the computer-aided design of longitudinal studies. *Frontiers in Psychology*, *6*, https://doi.org/10.3389/fpsyg.2015.00272

Brandmaier, A. M., von Oertzen, T., Ghisletta, P., Lindenberger, U., & Hertzog, C. (2018). Precision, reliability, and effect size of slope variance in latent growth curve models: Implications for statistical power analysis. *Frontiers in Psychology*, *9*, 1–16. https://doi.org/10.3389/fpsyg.2018.00294

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*, 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Duncan, S. C., Duncan, T. E., & Strycker, L. A. (2006). Alcohol use from ages 9 to 16: A cohort-sequential latent growth model. *Drug and Alcohol Dependence*, *81*(1), 71–81. https://doi.org/10.1016/j.drugalcdep.2005.06.001

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, *3*, 979–991. https://doi.org/10.1111/j.1751-9004.2009.00224.x

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242. https://doi.org/10.1037/h0044139

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin and Review*, *25*(1), 5–34. https://doi.org/10.3758/s1342

Ghisletta, P., McArdle, J. J., & Lindenberger, U. (2006). Longitudinal cognition-survival relations in old and very old age. *European Psychologist*, *11*(3), 204–223. https://doi.org/10.1027/1016-9040.11.3.204

Halpern, S., Karlawish, J., & Berlin, J. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association*, *288*(3), 358–362.https://doi.org/10.1001/jama.288.3.358

Hertzog, C., von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 541–563. https://doi.org/10.1080/10705510802338983

Hunter, W. G., & Hoff, M. (1967). Planning experiments to increase research efficiency. *Industrial and Engineering Chemistry*, *59*, 43–48. https://doi.org/10.1021/ie51402a010

Iddekinge, C. H. V., Ferris, G. R., Perrewé, P. L., Perryman, A. A., Blass, F. R., & Heetderks, T. D. (2009). Effects of selection and training on unit-level performance over time: A latent growth modeling approach. *Journal of Applied Psychology*, *94*, 829–843. https://doi.org/10.1037/a0014453

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. https://doi.org/10.2307/2291091

Kiken, L. G., Garland, E. L., Bluth, K., Palsson, O. S., & Gaylord, S. A. (2015). From a state to a trait: Trajectories of state mindfulness in meditation during intervention predict changes in trait mindfulness. *Personality and Individual Differences*, *81*, 41–46. https://doi.org/10.1016/j.paid.2014.12.044

Lau, M. A., Bishop, S. R., Segal, Z. V., Buis, T., Anderson, N. D., Carlson, L., . . . Devins, G. (2006). The Toronto Mindfulness Scale: Development and validation. *Journal of Clinical Psychology*, *62*(12), 1445–1467. https://doi.org/10.1002/jclp.20326

Lindenberger, U., & Ghisletta, P. (2009). Cognitive and sensory declines in old age: Gauging the evidence for a common cause. *Psychology and Aging*, *24*(1), 1–16. https://doi.org/10.1037/a0014986

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York, NY: Routledge.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science. *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*(2), 203–228. https://doi.org/10.1007/bf02294247

Rovine, M., & Molenaar, P. (1999). The covariance between level and shape in the latent growth curve model with estimated basis vector coefficients. *Methods of Psychological Research Online*, *3*, 95–107.

Saris, W., & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 181–204). Newbury Park, CA: Sage.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin and Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

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*, 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

von Oertzen, T. (2010). Power equivalence in structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 257–272. https://doi.org/10.1348/000711009X441021

von Oertzen, T., & Brandmaier, A. M. (2013). Optimal study design with identical power: An application of power equivalence to latent growth curve models. *Psychology and Aging*, *28*, 414–428. https://doi.org/10.1037/a0031844

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review*, *14*, 779–804. https://doi.org/10.3758/bf03194105

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York, NY: Springer. https://doi.org/10.1007/978-0-387-09612-4

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin and Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Received 16 April 2019; revised version received 20 September 2019

Appendix:

Computation of power equivalent models

Equations for the computation of power equivalent LGCMs following von Oertzen and Brandmaier (2013). For any original model with measurement occasions at t_{old}, we can construct a power equivalent model with $\tilde{n}$ measurement occasions at t_{new}.

$$t_{\text{new}} = \lambda \cdot \tilde{t}, \tag{1}$$

where

$$\begin{aligned}
\lambda &= \sqrt{\frac{\sigma_E^2}{\sigma_{\text{eff}}^2} \cdot \frac{\sigma_I^2 \tilde{n} + \sigma_E^2}{(\sigma_I^2 \tilde{n} + \sigma_E^2) \sum_{j=1}^{\tilde{n}} \tilde{t}_j^2 - \sigma_I^2 (\sum_{j=1}^{\tilde{n}} \tilde{t})^2}}, \\
\tilde{t}_j &= (j-1) \cdot \frac{\max(t_{\text{old}})}{\tilde{n} - 1}, \\
\sigma_{\text{eff}}^2 &= \frac{\sigma_E^2 (\sigma_I^2 \tilde{n} + \sigma_E^2)}{(\sigma_I^2 \tilde{n} + \sigma_E^2) \sum_{j=1}^{\tilde{n}} (\lambda \tilde{t}_j)^2 - \sigma_I^2 (\sum_{j=1}^{\tilde{n}} \tilde{t}_j)^2}.
\end{aligned} \tag{2}$$

The residual variance σ_E^2 and the intercept variance σ_I^2 are considered to be known and fixed and do not differ between models.

In our simulations, we used power equivalent models with 7, 5, and 3 measurement occasions. The power equivalent models with 5 and 3 measurement occasions were derived from the model with 7 measurement occasions at $t = 0, 1, ..., 6$.

In our application example, we used a design with 7 measurement occasions at $t = 0, 1, ..., 6$ as a starting point and derived power equivalent models with 3 and 10 measurement occasions from this model using the equations above. The variance of the intercept σ_I^2 and the error variance σ_E^2 were derived from Kiken, *et al.* (2015) and were set to $\sigma_I^2 = 43.6$ and $\sigma_E^2 = 21.45$.‘