How to Test Replication for Structural Equation Models

M.A.J. Zondervan-Zwijnenburg[1]

[1] Utrecht University

Author Note

This work was supported by the Consortium Individual Development (CID), which is funded through the Netherlands Organization for Scientific Research under grant: NWO Gravitation 024.001.003.

Correspondence concerning this article should be addressed to M.A.J. Zondervan-Zwijnenburg, Padualaan 14, 3584CH Utrecht. E-mail: m.a.j.zwijnenburg@uu.nl

Abstract

This paper introduces the prior predictive p-value as a manner to test replication in structural equation models. Using the replication of a piecewise latent growth model as a running example, the study explains the steps of the prior predictive p-value and illustrates them with R-code. The R-code included in the paper and the Supplementary R-script guides the reader through each analysis step. All steps to compute the prior predictive p-value are also incorporated in the `Replication` R-package. Finally, the study demonstrates how the replication of a more advanced structural equation model - a multilevel latent growth curve model - can be tested.

Keywords: Replication, Prior predictive p-value, Informative hypotheses

How to Test Replication for Structural Equation Models

The importance of conducting replication studies is increasingly recognized (Lindsay, 2015). Especially when an original study leads to remarkable and important findings, a new study may be conducted to see if the findings of the original study can be replicated. Several methods have been developed to test the replication of simple individual effect sizes. See for example, Anderson and Maxwell (2016), Harms (2018), Ly, Etz, Marsman, and Wagenmakers (2018) and Patil, Peng, and Leek (2016). To test the failure to replicate relevant findings obtained with ANOVA models, one can use the method presented in Zondervan-Zwijnenburg, Van de Schoot, and Hoijtink (2019). With this method, the replication of the ordering of the means, the difference between means, and the exact values of the means can be tested. No literature exists, however, that guides researchers in testing the replication of such relevant features in structural equation modeling (SEM), such as a piecewise growth curve model or a multilevel model. The current practice is to declare a factor structure replicated if it fits new data sufficiently, or to consider structural equation models replicated when the significance of paths and correlations is repeated in a second study (e.g., Carleton, Gosselin, & Asmundson, 2010; Stokes, Pogge, & Zaccario, 2013). A factor structure that repeatedly fits the data is indeed an indication of replication, just as repeated significance. These methods, however, do not formally test replication and are thus not able to reject replication. For example, failure to repeat significance can even occur with samples from the same population due to sampling variance and measurement error (Patil et al., 2016; Stanley & Spence, 2014). Hence, repeated model fit or significance do not lend themselves for conclusions about the (non-)replication of original results.

The current paper introduces the prior predictive p-value as a method to test replication in SEM. The prior predictive p-value provides an answer to the following replication research question: Does the new study fail to replicate relevant features of the original study? In answering this research question, the prior predictive p-value takes into account that the results of the new study may deviate from the original because of random

variation instead of meaningful differences. Furthermore, the current paper proposes to evaluate an informative hypothesis (Hoijtink, 2012) with the prior predictive p-value that contains the claims and relevant results of the original study. These claims can, for example, concern effect sizes, ordering of parameter magnitudes, or specific parameter values. In this manner, the prior predictive p-value focuses on the replication of relevant conclusions of the original study, and not on values of parameters that are nonessential to theory. Especially in SEM, where the model often encompasses many estimated parameters, the focus on the replication of relevant outcomes is an important feature.

The current paper will also explain step-by-step how replication of relevant SEM results can be tested with the prior predictive p-value (Box, 1980) in R (R Core Team, 2017) with the `Replication` package. In the background, the `Replication` package uses `lavaan` (Rosseel, 2012) en `blavaan` (Merkle & Rosseel, 2018) to analyze structural equation models. Readers should be familiar with the statistical software package R and the structural equation model that they want to analyze. Advanced statistical knowledge of, for example, Bayesian analyses is not required, but hands-on Bayesian knowledge can be very helpful (see, for example, Rupp, Dey, & Zumbo, 2004; Depaoli & Van de Schoot, 2017; Van de Schoot et al., 2013). As Supplementary Material, R-code and data are provided that can be used to execute each of the steps in this paper. The associated datasets are derived from the original datasets for illustrative purposes. Consequently, the results can differ from those in the manuscript, but the steps taken to arrive at the results are the same.

The next section describes the background and technical steps of the prior predictive p-value. The subsequent sections explain the steps to compute the prior predictive p-value in more detail with a piecewise latent growth model as a running example. Next, the prior predictive p-value is demonstrated for a multilevel latent growth curve model with predictors. The paper closes with a discussion and conclusion.

The Prior Predictive p-value

Zondervan-Zwijnenburg et al. (2019) introduced the idea that when we observe the results of an original study, it gives us expectations for future study results. This means that if we capture the original study results in a Bayesian posterior distribution, this posterior distribution holds our prior expectations for future data. The prior distribution contains a range of parameter values, with associated probabilities, that could all occur in future studies The prior predictive check (Box, 1980) uses the prior distribution and the statistical model to obtain a prior predictive distribution: a distribution with future datasets that can be observed given the prior (here: the original results). That is, given the prior expectations for parameters in future data, we simulate datasets. The next step is to compare the predicted datasets with the new observed dataset and compute a prior predictive p-value. To compare the datasets, we use what Box (1980) calls a 'relevant checking function'. The relevant checking function is a function that computes a relevant value with which the predicted data can be compared to the observed new data. Zondervan-Zwijnenburg et al. (2019) proposed to evaluate the misfit to a replication hypothesis H_0. The replication hypothesis H_0 is an informative hypothesis (Hoijtink, 2012) that is based on the claims and results of the original study. First, we compute the misfit to H_0 for each predicted dataset and for the observed new dataset. Next, we can compute the proportion of predicted datasets that scores more extreme with respect to H_0 than the new observed data. This is the prior predictive p-value. A small prior predictive p-value indicates that the new observed data is in the extreme end of results that we could obtain given the original study, considering H_0. A short technical explanation of each of the steps follows below, while the remainder of the study describes the steps in detail accompanied by R-code and empirical examples.

Step 1: The Prior Predictive Distribution

Let us denote the data by $\boldsymbol{y}_d$, where $d \in \{o, r, s\}$ with o for the original data, r for the new data, and s for predicted data. In the context of replication studies, we base the prior

for future data on the original study $\boldsymbol{y}_o$. That is, we sample from the posterior of the original study $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$.

$$g(\boldsymbol{\theta}_o|\boldsymbol{y}_o) \propto f(\boldsymbol{y}_o|\boldsymbol{\theta}_o)h(\boldsymbol{\theta}_o), \tag{1}$$

where $\boldsymbol{\theta}_o = \theta_{o1}, ..., \theta_{oJ}$ contains the J estimated model parameters for the original study, $f(\boldsymbol{y}_o|\boldsymbol{\theta}_o)$ is the likelihood of the original data, and $h(\boldsymbol{\theta}_o)$ is a prior distribution for the parameters of the original study. The prior $h(\boldsymbol{\theta}_o)$ should be specified such that the posterior is determined by the data. The posterior $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$ is our prior for future data $h(\boldsymbol{\theta}_s)$.

Using this prior distribution $h(\boldsymbol{\theta}_s)$ and the likelihood of the model at hand $f(\boldsymbol{y}_s|\boldsymbol{\theta}_s)$, the prior predictive distribution of new data can be determined, that is, the distribution of the data sets that are expected given the results of the original study:

$$\int f(\boldsymbol{y}_s|\boldsymbol{\theta}_s)h(\boldsymbol{\theta}_s)d\boldsymbol{\theta}_s = f(\boldsymbol{y}_s), \tag{2}$$

To obtain a discrete representation of the predictive distribution of the data $f(\boldsymbol{y}_s)$, we sample $b = 1, ..., B$ parameter vectors from $h(\boldsymbol{\theta}_s)$, and use them to simulate $t = 1, ..., T$ new datasets with sample size N_r. For simplicity, the collection of T samples from $f(\boldsymbol{y}_s)$ will be referred to as $f(\boldsymbol{y}_s)$. The section "Predicted Data" elaborates on the procedure to obtain $f(\boldsymbol{y}_s)$.

Step 2: The Replication Hypothesis H_0

To determine if the new study results significantly diverge from what we expect given the original study, we need to compare $\boldsymbol{y}_r$ to $f(\boldsymbol{y}_s)$. Many aspects of $\boldsymbol{y}_r$ to $f(\boldsymbol{y}_s)$ can be compared (e.g., mean values, maximum values, etc.), but we want to evaluate relevant features, which is what Box (1980) meant when he advised to use a 'relevant checking function'. In the context of replication, Zondervan-Zwijnenburg et al. (2019) propose to evaluate an informative replication hypothesis H_0 that is based on the results and conclusions of the original study. Both equality and inequality constraints among the parameters of the model at hand can be used to specify H_0, that is, H_0: $\boldsymbol{R\theta} > \boldsymbol{r}$ & $\boldsymbol{S\theta} = \boldsymbol{s}$

(Hoijtink, 2012; Silvapulle & Sen, 2005), where $\boldsymbol{R}$ and $\boldsymbol{S}$ are $K \times J$ restriction matrices, J denotes the number of estimated parameters, and K the number of restrictions in H_0, while $\boldsymbol{\theta}$ is the parameter vector of length J, and $\boldsymbol{r}$ and $\boldsymbol{s}$ are vectors of length K containing the constants in the replication hypothesis. The section "The Replication Hypothesis H_0" elaborates on the specification of H_0 with examples.

Step 3: The Prior Predictive p-value

Given H_0, we compute the test statistic D for the predicted and new data resulting in $f(D_s)$ and D_r. A useful and general operationalization of D is an approximate likelihood ratio test statistic of the constrained model in which $\boldsymbol{\theta}$ meets all restrictions given in H_0, and the unconstrained model with $\boldsymbol{\theta}$ estimated as usual to best fit the data at hand (Silvapulle & Sen, 2005, pp. 59–63):

$$\begin{aligned} D &= \ln \frac{f_u}{f_0} \\ &= (\ln f_u - \ln f_0), \end{aligned} \tag{3}$$

where

$$f_u = \max f(\boldsymbol{\theta}|\boldsymbol{y}_d), \tag{4}$$

that is, the unconstrained maximum likelihood for the parameters of interest, and

$$f_0 = \max_{\boldsymbol{\theta} \in H_0} f(\boldsymbol{\theta}|\boldsymbol{y}_d), \tag{5}$$

that is, the maximum likelihood for the parameters of interest under the constraints imposed by H_0.

It is not easy to obtain the maximum likelihood under the constraints imposed by H_0 for all statistical models. Hence, to compute f, we use a normal approximation of the density of the data: $\boldsymbol{\theta} \sim N(\boldsymbol{\theta}|\Sigma_{\boldsymbol{\theta}})$. In that case, we can use the `solve.QP` function from the `quadprog` R-package (Turlach & Weingessel, 2013), which finds the f_0 solution by approaching it as a quadratic programming problem. With a sufficiently large sample, it is

appropriate to use the variance-covariance matrix of the data $\Sigma_{\boldsymbol{\theta}}$, especially when the parameters in H_0 are unbounded, such as regression parameters and means. The approximate log-likelihood ratio may be less suited for small samples and bounded parameters in H_0 such as correlations and variances.

When we calculate D_s^t for each predicted dataset $\boldsymbol{y}_s^t$ given H_0, a discrete representation of the prior predictive distribution of the test statistic $f(D_{\boldsymbol{y}_s})$ is obtained. $f(D_{\boldsymbol{y}_s})$ is the distribution of the test statistic given for data that we expect given the original results. Finally, we can compute the prior predictive p-value:

$$P(D_{\boldsymbol{y_s}} \geq D_{\boldsymbol{y}_r}|H_0). \tag{6}$$

Given a predefined α, a significant prior predictive p-value makes us reject replication of the relevant findings in the original study: given the original results the new data obtain an extreme score with respect to H_0. Thus, considering H_0, the new data significantly deviate from the original results.

The next section illustrates each of the steps above in more detail with a piecewise latent growth model as a running example.

The Original Study

All replication efforts start with an original study. For example, Achterberg et al. (2017) evaluated the neural and behavioral correlates of social feedback and subsequent aggression in 74 7-10 year old children. The experiment consisted of 60 trials in which all children received 20 trials of positive, 20 trials of neutral, and 20 trials of negative feedback from an alleged unknown peer. In each trial, the children could respond to the feedback with a noise blast.

Below we take a look at the first six lines of the data $\boldsymbol{y}_o$, which is the object `y.o` in R. The data contains the average length of the noise blast in seconds per feedback condition (i.e., positive, neutral, negative).

```r
head(y.o) #head of data
```

```
##   positive  neutral negative
## 1 1.310263 1.724949 3.257900
## 2 1.603333 1.709088 2.791250
## 3 1.596444 1.769241 3.464211
## 4 2.600875 2.852826 3.180250
## 5 1.575500 1.849586 1.427650
## 6 1.942105 2.130500 3.500000
```

The statistical model in Achterberg et al. (2017) was a repeated measures model with each feedback condition as a repeated measure. To model all effects of interest (i.e., including differences between conditions), we use a piecewise latent growth model with fixed effects. Because the model specification is the same for all involved datasets (i.e., original, new, predicted), we will drop the subscript d in $\boldsymbol{y}_d$ when we discuss the model specifications. If we let $\boldsymbol{y}$ be a vector of length $p = 3$ with observed variables, the measurement model is given by:

$$\boldsymbol{y} = \boldsymbol{\nu} + \boldsymbol{\Lambda\eta} + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{\nu}$ is an item mean vector of length p, $\boldsymbol{\eta}$ is a vector with latent variables of length $q = 3$, $\boldsymbol{\Lambda}$ is a $p \times q$ matrix with factor loadings, and $\boldsymbol{\epsilon}$ is a vector with residuals for $\boldsymbol{y}$ of length p. $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Theta})$ where $\boldsymbol{\Theta}$ is a covariance matrix.

For the structural model, let $\boldsymbol{\alpha}$ be a vector with q latent means, and $\boldsymbol{\zeta}$ a vector with latent errors of length q:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\zeta}, \tag{8}$$

where $\boldsymbol{\zeta} \sim N(0, \boldsymbol{\Psi})$ with all elements of the covariance matrix $\boldsymbol{\Psi}$ equal to zero to have fixed (i.e., non-random) effects.

The piecewise latent growth model is modeled with an intercept (α_{i}) at the first measurement (i.e, the positive condition), a linear growth factor (α_{s1}) from the positive to the neutral condition and another linear growth factor (α_{s2}) from the neutral to the negative

condition. To estimate the latent factors, the elements in the item mean vector $\boldsymbol{\nu}$ are fixed at 0. Thus, our fixed effects piecewise latent growth model contains the following non-zero matrices:

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_{\text{positive}} \\ \boldsymbol{y}_{\text{neutral}} \\ \boldsymbol{y}_{\text{negative}} \end{bmatrix}, \boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \boldsymbol{\alpha} = \begin{bmatrix} \alpha_{\text{i}} \\ \alpha_{\text{s1}} \\ \alpha_{\text{s2}}, \end{bmatrix}, \boldsymbol{\Theta} = \begin{bmatrix} \Theta_{\text{positive}} & 0 & 0 \\ 0 & \Theta_{\text{neutral}} & 0 \\ 0 & 0 & \Theta_{\text{negative}} \end{bmatrix}.$$

As can be seen above, the estimated parameters are α_{i}, α_{s1}, α_{s2}, Θ_{positive}, Θ_{neutral}, and Θ_{negative}.

The `lavaan` model syntax is provided in Appendix A and in the Supplementary R-code. In the model, we can also compute the effect sizes between the different measurements. To do that, we divide the effect of interest by the pooled standard deviation. That is: $d_{\alpha_{\text{s1}}} = \frac{\alpha_{\text{s1}}}{(\sqrt{\Theta_{\text{positive}}+\Theta_{\text{neutral}}})/2}$ (i.e., the standardized difference between the positive and neutral condition), $d_{\alpha_{\text{s2}}} = \frac{\alpha_{\text{s2}}}{(\sqrt{\Theta_{\text{neutral}}+\Theta_{\text{negative}}})/2}$ (i.e., the standardized difference between the neutral and negative condition), and $d_{\alpha_{\text{s1}}+\alpha_{\text{s2}}} = \frac{\alpha_{\text{s1}}+\alpha_{\text{s2}}}{(\sqrt{\Theta_{\text{positive}}+\Theta_{\text{negative}}})/2}$ (i.e., the standardized difference between the positive and negative condition). The pooled standard deviations and effect sizes are not estimated, instead they are derived from the estimated parameters.

If we store the model syntax in `model.A`, we can run the piecewise latent growth model in the R-package `lavaan` as shown below.

```
library(lavaan)
fit.o <- sem(model=model.A,data=y.o)
```

The resulting latent means, residual variances, and effect sizes are:

```
##            id  est   se p-value
## a_i        10 1.58 0.10   0.000
## a_s1       11 0.39 0.12   0.001
## a_s2       12 0.88 0.11   0.000
## e_positive 13 0.71 0.12   0.000
## e_neutral  14 0.41 0.07   0.000
```

```
## e_negative 15 0.45 0.07   0.000
## d_s1       28 0.53 0.17   0.002
## d_s2       29 1.35 0.18   0.000
## d_s1+s2    30 1.68 0.19   0.000
```

The column `id` shows the parameter identification value assigned by `lavaan`, the column `est` shows the estimate, the column `se` shows the standard error, and the column `p-value` contains the p-value. It is important to note the standard error. Standard errors inform us about the accuracy of a parameter. The larger the standard error, the wider the confidence interval for the parameter, and the more future findings will be in line with the original finding. For a useful replication test, the original study needs to produce specific results and conclusions. Original studies that are suitable for replication testing contain statistically significant findings and effect sizes that are at least of a medium size. Both indicators are present in Achterberg et al. (2017).

Given that we have an original study that is suitable for replication testing, we can take two steps to compute the prior predictive p-value: Step 1: The Prior Predictive Distribution, and Step 2: The Replication Hypothesis H_0. We will first continue with Step 1, in which we predict what new datasets can look like given the current original results.

The Prior Predictive Distribution

The prior predictive distribution is a distribution of predicted datasets given the model and prior distribution. If we expect the original study to replicate, then the original study contains prior information for future datasets. Following this line of reasoning, we let the results of the original study determine the prior predictive distribution. The results of the original study are captured in the posterior distribution that results from a Bayesian analysis of the original data. Thus, the posterior $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$ is our prior for expected data $h(\boldsymbol{\theta}_s)$. In this manner, we base the prior predictive distribution on the original results.

The posterior distribution is calculated through an iterative process (e.g., Markov

chain Monte Carlo) in which each iteration results in a set of parameter values. To begin the iterative procedure, starting values are used. Over the course of iterations, the impact of the starting values on the results diminishes and is expected to disappear. To remove the impact of the starting values, the first couple of thousands of iterations are regarded as burn-in iterations and they are not included in the posterior distribution.

One practical advantage of the posterior distribution is that we can easily sample model parameter values from it. For example, for Achterberg et al. (2017), we can take a sample from the posterior distribution of the parameters in the piecewise latent growth model. This sample contains values for α_{i}, α_{s1}, α_{s2}, Θ_{positive}, Θ_{neutral}, and Θ_{negative}. The parameter values are our prior information to predict (i.e., simulate) future data under the same statistical model.

The R-package `Replication` can (1) obtain the posterior distribution for the original data, (1) draw samples from the posterior distribution, and (3) simulate future data within one function: `ppc.step1`. The remainder of this section, however, elaborates step-by-step what happens in this function and why. The data and code to reconstruct all output are provided in the Supplementary Materials.

As a first step, we run a Bayesian analysis on the Achterberg et al. (2017) data with the R-package `blavaan` (Merkle & Rosseel, 2018) and the default prior distributions (see Appendix B), and the default number of 5,000 burn-in and 10,000 post burn-in iterations. An in-depth guide on how to run and evaluate a Bayesian analysis is Depaoli and Van de Schoot (2017).

```
library(blavaan)

b.fit <- bsem(model=model.A,data=y.o)
```

Here, `model.A` is the `lavaan` syntax for the piecewise latent growth model as described in the previous section and provided in Appendix A.

An essential step in a Bayesian analysis is to evaluate convergence of the estimation process (Depaoli & Van de Schoot, 2017). Therefore, we request traceplots.

```
plot(b.fit,plot.type="trace")
```

Traceplots are generated for each estimated parameter. The traceplots show the parameter value (y-axis) for each iteration (x-axis) in the iterative process. Separate lines in the plot, represent separate estimation "chains". If the estimation process went well, traceplots have a stable mean and variance and separate chains mix well. We then call the analysis 'converged'. Figure 1 shows the traceplots for four of the estimated parameters in the model of Achterberg et al. (2017): α_{i} denoted by the syntax `i~1`, α_{s1} denoted by the syntax `s1~1`, α_{s2} denoted by the syntax `s2~1`, and Θ_{positive} denoted by the syntax `positive~~positive`.

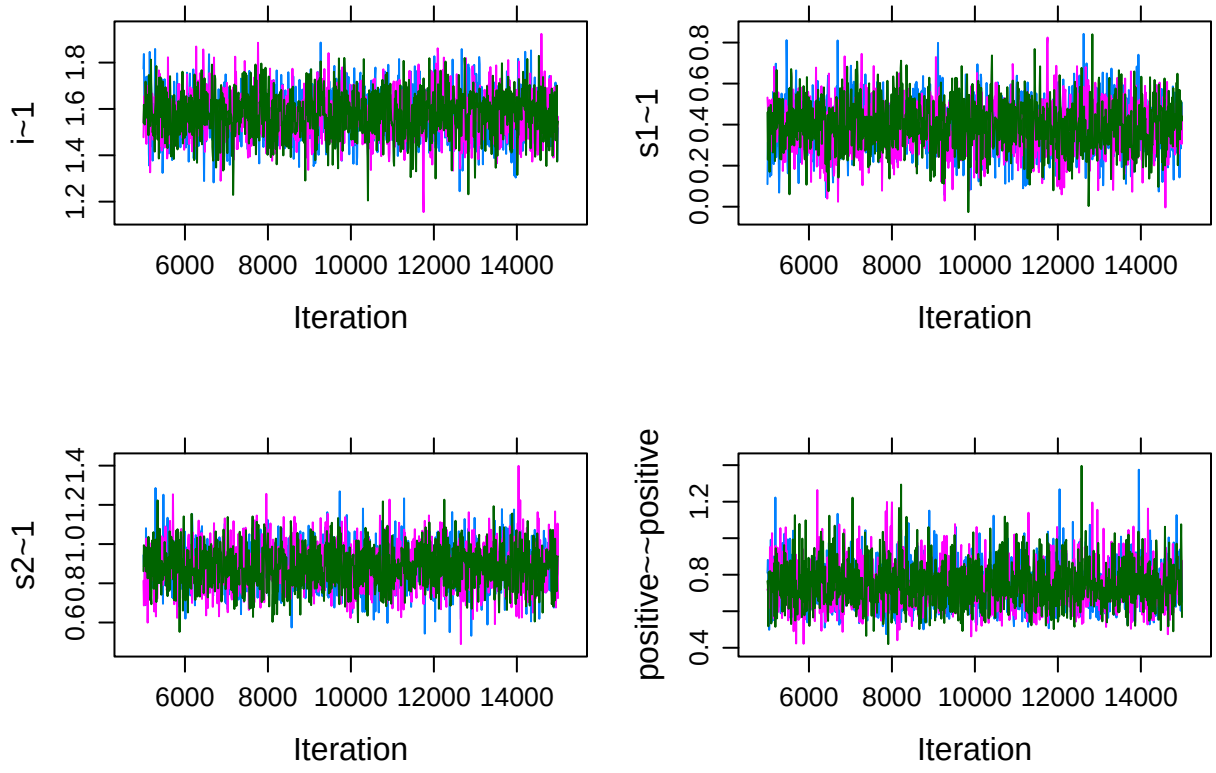

Figure 1. Traceplots.

Indeed, the means and variances for the iterations in these plots (i.e., iteration 5,000-15,000) look sufficiently stable. If convergence has not been achieved, we can, for example, increase the number of iterations or evaluate whether prior distributions other than

the default `blavaan` priors for the parameters (see Appendix B) may suit the analysis better. All in all, we consider the analysis successful, and we can use the results to obtain predicted datasets conditional on the original results.

To illustrate the concept of a posterior distribution, we also show the histogram that depicts the samples from the posterior for α_i in Figure 2. Each bar in the histogram counts how often the estimation process resulted in the associated range of values. The more iterations are used in the analysis, the smoother the histogram will look.

```
plot(b.fit,plot.type="hist",pars=1)
```

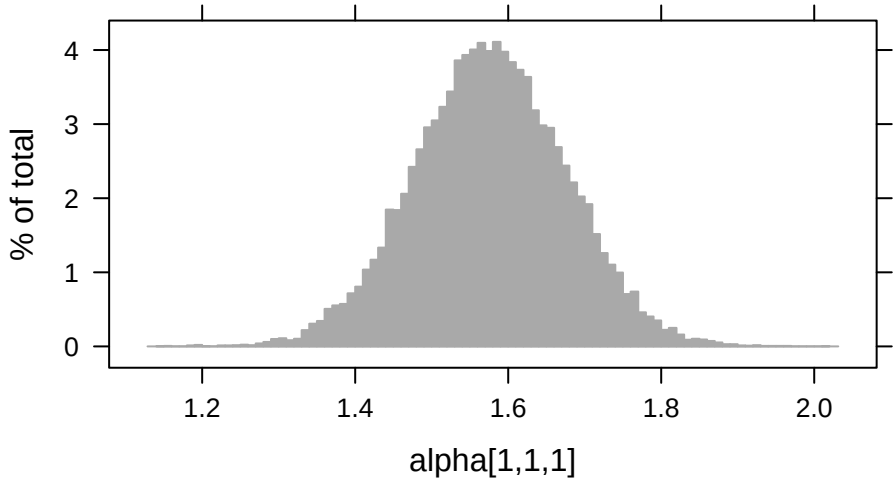

Figure 2. Samples from the posterior for α_i.

To obtain predicted datasets, we make use of the fact that the Bayesian analysis provided us with a set of parameter values in each iteration. Below we request the parameter values (i.e., latent means and residual variances) from the first post burn-in iteration of the Achterberg et al. (2017) analysis.

```
posterior <- blavInspect(b.fit,"mcmc")
posterior[[1]][1,1:6]
```

```
## alpha[1,1,1] alpha[2,1,1] alpha[3,1,1]
##         1.77         0.11         0.94
```

```
## theta[1,1,1] theta[2,2,1] theta[3,3,1]
##         0.81         0.51         0.56
```

The parameter values in this posterior sample differ from the `blavaan` summary results presented before, because the summary presents the mean results over all iterations, and here we see a single iteration.

We can feed each model parameter value from a selected posterior sample to simulation software, such as the `simulateData` function in the R-package `lavaan` (Rosseel, 2012). This function then simulates a dataset under the imposed model with the drawn set of posterior parameter values as population parameter input. We request that this future dataset has the sample size of the new dataset, in order to get the prior predictive distribution of data sets with the same size as the new data set based on the results of the original study. As stated before, our simulation results in a discrete representation of the prior predictive distribution. To produce a proper representation of the prior predictive distribution, we sample many times from the posterior distribution of the original data, for example, 5,000, and predict a dataset for each of those samples.

The function `ppc.step1` in the `Replication` package takes (1) the statistical model of interest, (2) the original dataset, and (3) the sample size of the new dataset. By default, the function will use 2 chains, 1,000 model adaptation iterations, 5,000 burn-in samples, 5,000 post burn-in samples and default priors to obtain the posterior distribution of the original data. Furthermore, the function will by default simulate 5,000 datasets for the prior predictive distribution. The default settings can be adjusted, and optional commands can be added as well. For example, the user can also choose to let the Bayesian software continue to iterate until it determines that convergence is achieved with `convergence = "auto"`. The function can also be used with missing data (see the Section “Missing Data”). Other `lavaan` or `blavaan` modelling commands (e.g., multiple group analysis, type of estimator) can be added as well, but these are optional. Run `?ppc.step1` for a full overview of options and their descriptions. Run `?ppc.step1` for a full overview of options and their descriptions.

Here we load the `Replication` package and apply the function to the data of Achterberg et al. (2017) with the required arguments only, using the default settings for the remaining arguments.

```
library(Replication)
step1.A <- ppc.step1(y.o=y.o,model=model.A,n.r=nrow(y.r))
```

As a result, the `ppc.step1` function automatically (1) runs the Bayesian analysis, (2) generates the traceplots and summary of the results for verification, (3) draws 5,000 sets of parameters from the posterior distribution, and (4) simulates 5,000 datasets to represent the prior predictive distribution. The 5,000 datasets that we obtained are datasets that can occur given the results of the original study with the sample size of the new study. We can take a look at the top of the first predicted dataset:

```
head(step1.A$y.s[[1]])
```

```
##     positive  neutral negative
## 1  1.0356699 1.348873 2.733568
## 2  2.1204153 1.303761 2.782484
## 3 -0.6949553 2.710005 4.242003
## 4  0.5501848 1.616024 2.833882
## 5  2.3839132 2.324915 3.364297
## 6  1.3930865 2.459061 2.532939
```

As you can see, the values are not equal to those in the original dataset, because we do not need to expect exactly the same values for replications of the original study. The predicted observations, however, relate to the original study in the sense that they are predicted based on parameter values for the original dataset.

To compare the predicted data to the observed new data, we need to determine what relevant features are to compare the datasets by. These features will be captured in the replication hypothesis H_0. Hence, the second step to compute the prior predictive p-value is

to define the replication hypothesis H_0, which is further explained in the next section.

The Replication Hypothesis H_0

The findings of the original study can be summarized in an informative hypothesis H_0 (Hoijtink, 2012; Silvapulle & Sen, 2005; Zondervan-Zwijnenburg et al., 2019). An informative hypothesis is a hypothesis that contains information about model parameters. By means of constraints, the informative hypothesis limits the values that the parameter is allowed to take on. Types of constraints are: range constraints, order constraints, and equality constraints (Silvapulle & Sen, 2005).

Consider the case of Achterberg et al. (2017), where the statistical model is a piecewise latent growth model with the estimated parameters α_{i}, α_{s1}, α_{s2}, $\Theta_{\text{positive}}, \Theta_{\text{neutral}}$, and Θ_{negative}. In the original study we found: $\alpha_{\text{i}} = 1.58$, $\alpha_{\text{s1}} = 0.39$, $\alpha_{\text{s2}} = 0.88$, $\Theta_{\text{positive}} = 0.71$, $\Theta_{\text{neutral}} = 0.41$, and $\Theta_{\text{negative}} = 0.45$. Additionally, $d_{\alpha_{\text{s1}}} = 0.53$, $d_{\alpha_{\text{s2}}} = 1.35$, and $d_{\alpha_{\text{s1+s2}}} = 1.68$.

An informative hypothesis contains a range constraint when it specifies the range of values that the parameters are in. For example, H_0: $\alpha_{\text{i}} > 1.5$, $\alpha_{\text{s1}} > 0$, $\alpha_{\text{s2}} > 0.5$. An order constraint, on the other hand, specifies how certain parameters relate to each other, for example, H_0: $\alpha_{\text{s1}} < \alpha_{\text{s2}}$. Alternatively, an equality constraint can, for example, have the following forms: $\alpha_{\text{s1}} = \alpha_{\text{s2}}$, or $\alpha_{\text{i}} = 1.58$. Note that these examples do not include information on $\Theta_{\text{positive}}, \Theta_{\text{neutral}}$, and Θ_{negative}. The reason that the residuals are not included in H_0 is that the original study makes no claims about these parameters. Hence, we do not want to put a restriction on them.

The content of the replication hypothesis H_0 depends on the claims and results of the original study. For example, Achterberg et al. (2017) state:

> "The combined effect for the difference between positive and neutral was medium in size ... The difference between neutral and negative feedback showed a large combined effect size ... The difference between positive and negative

feedback also showed a large combined effect size ..." (p. 111)

Based on these claims, we want to test the replication of these effect sizes. For effect sizes, Zondervan-Zwijnenburg et al. (2019) recommend to use the lower limits of Cohen's effect size categories (i.e., .20 for a small effect, .50 for a medium effect, and .80 for a large effect) as a lower limit for replication in the replication hypothesis. Thus, in the case of Achterberg et al. (2017) we specify the following range hypothesis, H_0: $d_{\alpha_{s1}} > .50$, $d_{\alpha_{s2}} > .80$, and $d_{\alpha_{s1}+\alpha_{s2}} > .80$.

If the claims by the original study do not concern effect sizes, but rather highlight the significance of certain parameters, the user needs to determine the reasonable lower limit for replication. As an example, consider that we have two statistically significant parameters of interest: $\alpha_{s1} = 0.39$ and $\alpha_{s2} = 0.88$. Some options that we have for H_0 are:

1. H_0: $\alpha_{s1} > 0$, $\alpha_{s2} > 0$
2. H_0: $\alpha_{s1} > 0.30$, $\alpha_{s2} > 0.50$
3. H_0: $\alpha_{s1} > 0.39$,$\alpha_{s2} > 0.88$

The main criterion is that the lower limit in H_0 needs to stay close to the original study and its theory. When the content of the informative hypothesis H_0 is determined, the hypothesis needs to be formalized into a more technical format that can be used by software such as R (R Core Team, 2017).

To include the hypotheses in the `Replication` package, we specify the hypothesis within quotes with the `plabels` given in the parameter table resulting form `ppc.step1` as variable names and `&` to separate constraints within the hypothesis. If H_0 concerns effect sizes, a vector `s.i` includes the `id` values of the (pooled) standard deviations in the summary table produced by `ppc.step1` by which the parameters of interest should be standardized.

The replication hypothesis for Achterberg et al. (2017) is H_0: $d_{\alpha_{s1}} > .50$, $d_{\alpha_{s2}} > .80$, and $d_{\alpha_{s1}+\alpha_{s2}} > .80$. To prepare this hypothesis as input for the `Replication` package, we look at the parameter table resulting from `step1.A` and identify the `plabels` for the coefficients of interest. Furthermore, we identify the blavaan id's of the pooled standard

deviations.

```
#have a look at the parameter table and identify latent slope factors
step1.A$pT
#s.i identify id of pooled s coefficients by their defined labels
pT <- step1.A$pT
s.i <-  c(pT$id[which(pT$lhs=="s12")],pT$id[which(pT$lhs=="s23")],
          pT$id[which(pT$lhs=="s13")])
```

We find that the `plabel` for α_{s1} = `.p11.`, and for α_{s1} = `.p12.`. Thus, the hypothesis is `".p11.>.50 & .p12.>.80 & .p11.+.p12.>.80"` with `s.i=s.i`.

To recap briefly, we can compose an informative replication hypothesis H_0 that captures the main findings of the original study. The replication hypothesis is the key element to compute the test statistic D that we use to compare the new observed dataset to the predicted datasets. Before we compute the test statistic, however, we will discuss the new study in the next section.

The New Study

Next to the predicted data that the `Replication` package creates, we have the observed new dataset. The new dataset is the result of a replication effort. Just as for original studies, it is important that the new study has a substantial sample size that yields sufficient power to test the model at hand. Muthén and Muthén (2002) explains how a Monte Carlo study can be used to estimate the required sample size. In the context of replication, Simonsohn (2015) recommends that the sample size for the new study is 2.5 times the original sample size.

The new study in this example is Achterberg, Duijvenvoorde, Van der Meulen, Bakermans-Kranenburg, and Crone (2018). The behavioral task in Achterberg et al. (2018) is a direct replication of the behavioral task in Achterberg et al. (2017) with 509 participants, which is almost 7 times the original sample size. For this data we want to test

whether Achterberg et al. (2018) deviates more from Achterberg et al. (2017) with respect to H_0: $d_{\alpha_{s1}} > .50$, $d_{\alpha_{s2}} > .80$, and $d_{\alpha_{s1}+\alpha_{s2}} > .80$ than expected by chance. How we can conduct this test is described in the next section.

Computing the Prior Predictive p-Value

We now have (1) obtained the prior predictive distribution and (2) set the replication hypothesis H_0. When we confront the elements obtained in step 1 and 2 with the new data, we can obtain the prior predictive p-value in the third and final step of this procedure. We want to compare whether the new dataset is similar to the predicted datasets considering the replication hypothesis. To make the comparison, we compute for each dataset the approximate likelihood ratio statistic D as presented in Equation 3. The statistic D reflects how much the dataset deviates from the replication hypothesis H_0. If $D = 0$, there is no difference between the parameters estimated under the unconstrained hypothesis H_u and the ones that are fitted under the constraints of H_0. In other words, if $D = 0$ the unconstrained parameter estimates fit H_0 perfectly.

When the new dataset and all predicted datasets have a score D, we can compare the new observed dataset to the predicted datasets. We can compute the proportion of predicted datasets that obtains the same, or a larger D score than the observed new dataset. This is the prior predictive p-value (See also Equation 6). The smaller the prior predictive p-value, the more extreme the new observed dataset scores with respect to H_0 as compared to predicted datasets given the original data. If the prior predictive p-value is smaller than a preset Type I error rate α, we can reject replication of the original study considering H_0. The prior predictive check cannot 'prove' replication, but an indication of a replication is provided if replication cannot be rejected while the test had sufficient statistical power. For simple statistical models (e.g., univariate models), it can be possible to compute the statistical power to reject replication (Zondervan-Zwijnenburg et al., 2019). Generally, it is recommended to make use of well-powered original and new studies, where the new study is

preferably 2.5 times the size of the original study (Simonsohn, 2015).

To obtain the prior predictive p-value, we make use of the function `ppc.step2step3` of the `Replication` package. The function `ppc.step2step3` first computes D for each dataset (i.e., predicted data and observed new data), and then applies Equation 6, which yields the prior predictive p-value. We provide the function with: (1) the results of `ppc.step1`, (2) the new data, (3) the statistical model, (4) H_0, and (5) the vector `s.i` including the `id` values of the (pooled) standard deviations in the summary table produced by `ppc.step1` by which the parameters of interest should be standardized. Other `lavaan` or `blavaan` modelling commands (e.g., multiple group analysis, type of estimator) can be added as well, but these are optional. Run `?ppc.step2step3` for all options. The `R` code for the running example is:

```
H0 <- ".p11.>.50 & .p12.>.80 & .p11.+.p12.>.80"
step23.A <- ppc.step2step3(step1=step1.A,y.r=y.r,model=model.A,
                           H0=H0,s.i=s.i)
```

The resulting D for the new data and prior predictive p-value are requested as follows.

```
step23.A$llratio.r #D in new data
step23.A$`p-value` #p-value
```

Figure 3 shows a histogram of D for $\boldsymbol{y}_s$. A thick black line at $D = 0$ on the x-axis indicates that more than 2,500 of the 5,000 predicted datasets perfectly matched H_0. Larger values of D also occur in the predicted data. This may seem surprising to some, because the predictive distribution was based on the original study that also produced H_0. This deviance between H_0 and the predicted data can occur as a result of random variation.

For Achterberg et al. (2018) $D = 0$, as is also illustrated by the red vertical line in Figure 3. This means that the new data perfectly follows the replication hypothesis H_0. As a result, the prior predictive p-value is 1.000. Thus, all predicted datasets have the same or a more extreme deviation from H_0. The prior predictive p-value shows that we cannot reject replication of the original study results. Since $D = 0$ and the prior predictive $p = 1$, we can

even state that the new study replicates the replication hypothesis generated by Achterberg et al. (2017).

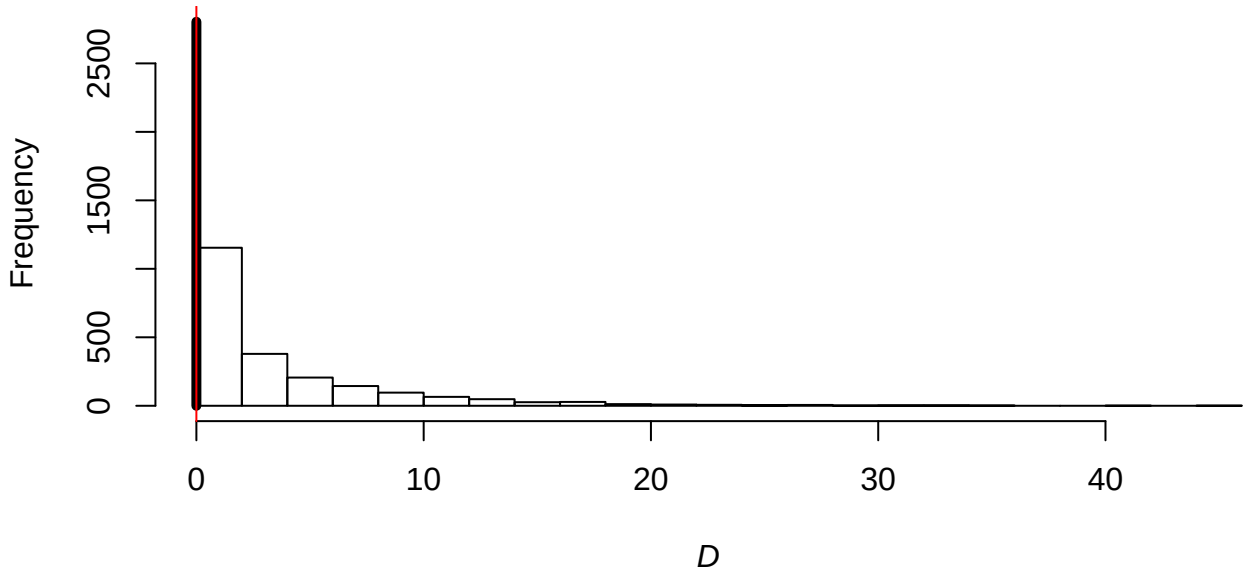

Figure 3. Histogram of predicted D for the replication of Achterberg et al. (2017) with the observed D for Achterberg et al. (2018) indicated by the red line

Generally, the prior predictive p-value tells us whether the new study significantly deviates from what we would expect based on the original findings, considering H_0. If $p > .05$ but < 1, the result is not perfectly in line with the original findings and we can only conclude that we cannot reject replication of the original study. If the original study results concerned large effects based on a sufficient sample size and if the new study sample size was sufficient as well, we did not prove replication, but replication is a likely interpretation of results. If the original study results were vague and sample sizes were insufficient, we should consider a lack of power as an alternative explanation for not rejecting replication of the original study results.

In sum, (1) we have predicted datasets given the original findings, and (2) using the replication hypothesis (3) we have compared the new observed dataset to the predicted datasets by their deviance from the replication hypothesis. The result is a prior predictive p-value that indicates whether we reject replication of the original study considering its

relevant findings. We have accomplished this using only two functions of the `Replication` R-package: `ppc.step1` and `ppc.step2step3`. The steps to compute the prior predictive p-value can be applied to other studies, models, datasets, and hypotheses as well. The next section illustrates how replication can be tested in three steps for a multilevel latent growth model with predictors.

An Example of a Multilevel Longitudinal Growth Curve Model

Bakker, Van der Heijden, Van Son, and Van Loey (2013) examined traumatic stress reactions in couples after a burn event to their preschool child (0-4 years). The couples, representing 190 children, reported four times in 18 months on their intrusion and avoidance symptoms. We will focus on the intrusion results. Bakker et al. (2013) used a three-level model (time in parents in couples) to analyze the development and predictors of intrusion. The model including the cross-level regressions is depicted in Figure 4. The top of Figure 4 shows the time level with three repeated measurements of intrusion: int0, int3, in12 and int18. Three latent factors capture the intercept of intrusion at the first measurement in i, the linear growth rate per month at the first measurement in s, and the quadratic factor in q. The second level is the parent level with predictors measured in fathers and mothers: anger, guilt, parent gender, and feelings of threat. The third level is the couple level with predictors for the parent couple: gender of the child, age of the child, burn size, and location of the burn event (i.e., inside or outside the home). The intercept of intrusion is regressed on all parent and couple predictors (i.e., β_{anger}, β_{guilt}, β_{genderP}, β_{threat}). The linear slope of intrusion is regressed on anger and parent gender (i.e., $\beta_{\text{anger*s}}$, $\beta_{\text{genderP*s}}$). Egberts, Schoot, Geenen, and Van Loey (2017) repeated the study of Bakker et al. (2013) with parents of school-aged children (8-18 years) that were subject to a burn event. That is, in their study 111 mothers and 91 fathers of 108 children reported four times in 18 months on their intrusion and symptoms.

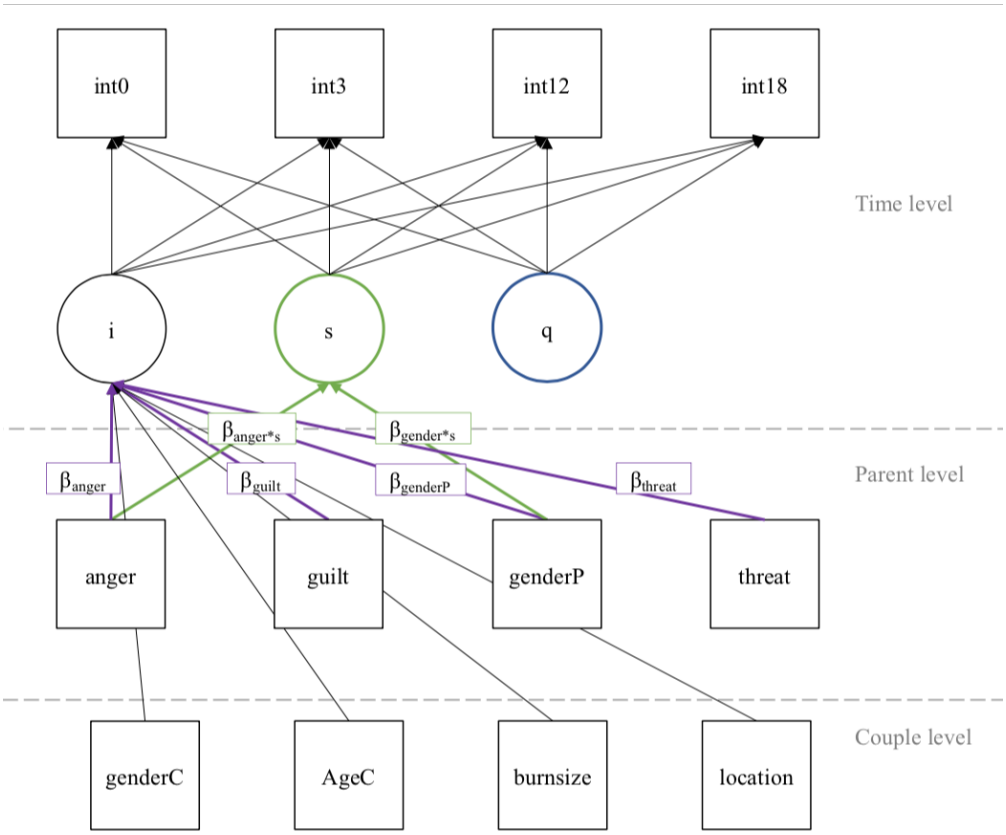

Figure 4. Multilevel model as evaluated in Bakker et al. 2013.

The Prior Predictive Distribution

First, the multilevel model was rewritten at the first level for wide format data (see Supplementary R-code), because `blavaan` does not include a cluster function yet. Because the multilevel longitudinal growth-curve model is relatively complex, we first conducted a preliminary analysis with automatic convergence settings, which indicated that about 25,000 post burn-in iterations would result in convergence for all parameters. Hence, we set the number of post burn-in iterations in the `ppc.step1` function to 25,000 to obtain the predicted data.

```
step1.B.M <- ppc.step1(y.o=y.o,model=model.B,n.r=n.r,nchains=3,nsample=25000)
```

As a result of running `ppc.step1`, we obtained trace plots for each parameter, a parameter table with information about the parameters such as estimates, and (the default

number of) 5,000 predicted datasets that represent future data given the original findings. The trace plots showed acceptable convergence. Hence, the next step was to specify the hypothesis of interest H_0 with which we could compare the predicted datasets to the observed new dataset.

The Replication Hypothesis H_0

With respect to intrusion, Bakker et al. (2013) drew conclusions about the eight bold and colored parameters in Figure 4:

> "Mothers had higher scores [positive β_{genderP}] ... A general decline in intrusion was observed in all parents [negative latent factor s], but a small quadratic term for time indicated that this decrease in symptoms was not strictly linear [small positive latent factor q] ... Parents within couples did not have the same course of symptoms over time ("random slopes"). For symptoms of intrusion, the difference between mothers and fathers became smaller over time [negative $\beta_{\text{genderP*s}}$]." (p. 1079)

> "For intrusion, the final model with explanatory variables showed that apart from parent gender, perceived threat to the child's life [positive β_{threat}]... and parental feelings of guilt [positive β_{guilt}]... affected the level of symptoms throughout the entire study period ... For early feelings of anger, the results showed an initial influence on symptoms of intrusion [positive β_{anger}]..., but this influence diminished as time passed [negative $\beta_{\text{anger*s}}$]..." (p. 1080)

> All mentioned findings were observed with one-sided p-values smaller than .01.

From the parameter table obtained with `ppc.step1`, we derived the replication parameter estimates for H_0. In translating these findings into a replication hypothesis, an expert on the subject judged whether the same estimates for parameters in H_0 could be expected for the older children in the new study. According to the expert, intercepts for predictors may change, but child age is not a reason to expect different values for the latent

time variables and regression parameters of interest. Hence, H_0: s < -0.67, q > 0.03, $\beta_{\text{genderP}} > 4.68$, $\beta_{\text{guilt}} > 0.77$, $\beta_{\text{anger}} > 1.35$, $\beta_{\text{threat}} > 3.69$, $\beta_{\text{genderP*s}} < -0.14$, $\beta_{\text{anger*s}} < -0.07$. Again, we identify the `plables` and include them in an object $H0$ (see Supplementary Materials for an automatized procedure).

The Prior Predictive p-Value

With the prior predictive p-value we can check the agreement with H_0 in Egberts et al. (2017). If we separately analyze Egberts et al. (2017). We obtain the following results: s $= -0.70$, q $= 0.03$, $\beta_{\text{genderP}} = 5.59$, $\beta_{\text{guilt}} = 0.96$, $\beta_{\text{anger}} = 0.64$, $\beta_{\text{threat}} = 2.15$, $\beta_{\text{genderP*s}} = -0.15$, $\beta_{\text{anger*s}} = 0.00$. We can see that the results are not perfectly in line with H_0, but the question is: do they deviate more than what we would expect based on random variation?

To answer this question, we provide the `pcc.step2step3` with (1) the results of step1 stored in `step1.B.M`, (2) the new data `y.r`, and (3) the hypothesis stored in `H0`.

```
step23.B <- ppc.step2step3(step1=step1.B.M,y.r=y.r,
                           model=model.B,H0=H0)
```

For Egberts et al. (2017) $D = 11.18$, and the prior predictive p-value is 0.011 (See also Figure 5). The new data by Egberts et al. (2017) scored in the extreme 1.1% of the predicted data with respect to the replication hypothesis. Hence, we reject the replication of H_0 by Egberts et al. (2017).

All in all, the example above showed how replication for a structural equation model can be tested in a few steps using the `Replication` package to compute the prior predictive p-value. So far, however, an original and new dataset in which missing values were imputed once, which is not proper for inferences. Since missing data is common in social science research, we comment on this topic in the next section and suggest how it could be dealt with. Our proposal, however, is not ideal yet for missing data in $\boldsymbol{y}_r$.

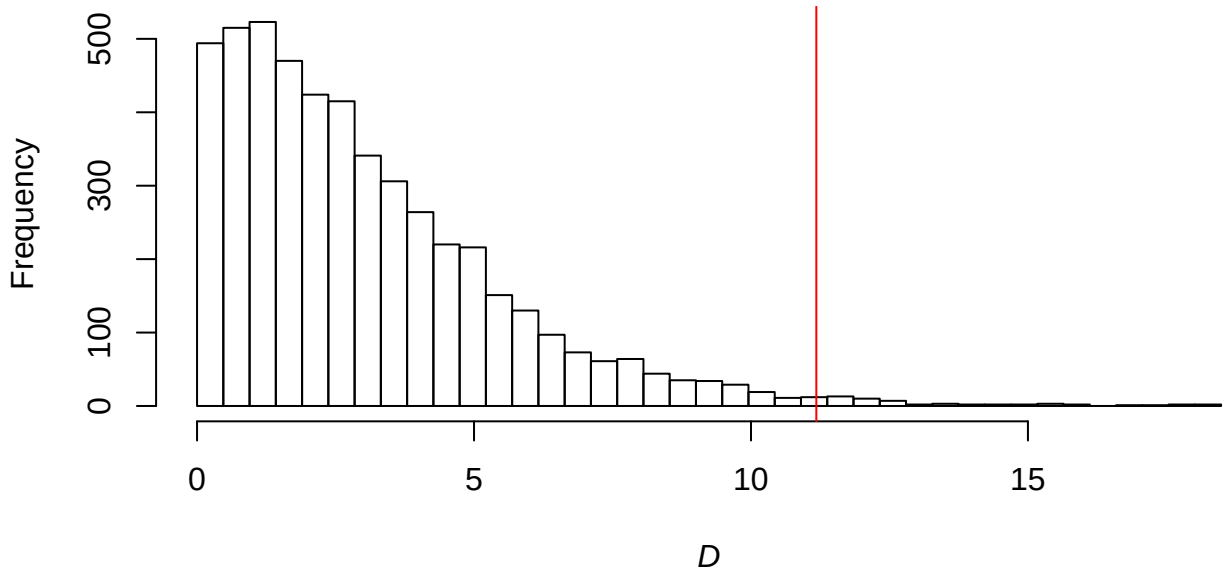

Figure 5. Histogram of predicted D for the replication of Bakker et al. (2013) with the observed D for Egberts et al. (2017) indicated by the red line.

Missing Data

Missing data in $\boldsymbol{y}_o$ can be resolved by applying multiple imputation on $\boldsymbol{y}_o$. A basic example of imputation with `mice` (Van Buuren & Groothuis-Oudshoorn, 2011) on the wide format data is shown below:

```
library(mice)
#choose predictor variables; exclude ID and collinear gender variable
pred.1 <- quickpred(p1.w,mincor = 0.25,exclude=c("fam","ParentG.1"))
#impute the data
imp.B  <- mice(p1.w, maxit=25, m=10, predictorMatrix=pred.1)
#evaluate the imputation
imp.B$loggedEvents; plot(imp.B)
```

Here, `imp.B` is the object with imputed datasets. Subsequently, we can compute the posterior distribution for each imputed dataset and combine those posterior samples (Gelman et al., 2013, pp. 451–452). The function `ppc.step1` will do this automatically if an

object imputed with `mice` is included. The input for the `y.o` argument is then ignored. Hence, our input with missing data in `y.o` is:

```
step1.B.M_mis <- ppc.step1(y.o=y.o,model=Model.B.mis,n.r=n.r,
                           imp=imp.B,nsample=25000)
```

where `Model.B.mis` is the statistical model of Bakker et al. (2013) without the quadratic factor to facilitate estimation.

Missing data in $\boldsymbol{y}_r$ poses a problem for the prior predictive p-value, because the predicted data needs to be comparable to the new data, and thus needs to have the same sample size and no missing data. To circumvent this issue, we propose to compare complete datasets by applying multiple imputation on $\boldsymbol{y}_r$ and comparing all M complete datasets to the predicted data $f(\boldsymbol{y}_s)$, which has the same sample size. As a result, we obtain M prior predictive p-values. If all prior predictive p-values are non-significant while the sample size in the new study was sufficient, it appears that the new study does not deviate more from H_0 than we would expect given the original results. The more prior predictive p-values are significant, the more doubt we have that the new study does not replicate the original results as captured by H_0.

In the `Replication package`, we can obtain the prior predictive p-values for replicated data in two steps. First, we run the `pcc.step2step3` as usual, but now with `y.r = NULL`. Consequently, the function will only evaluate the replication hypothesis H_0 for the predicted datasets $\boldsymbol{y}_s$ and not for $\boldsymbol{y}_r$.

```
step23.B.M <- ppc.step2step3(step1=step1.B.M_mis,y.r=NULL,
                             model=Model.B.mis,H0=H0)
```

Second, we evaluate H_0 for each imputed new dataset to obtain a distribution of D scores and prior predictive p-values. To obtain D for each imputed dataset, we use the function `llratio.imp`. We provide `llratio.imp` with the results of `ppc.step2step3`, the imputed `mice` object, and the model.

```
robust <- llratio.imp(step2step3=step23.B.M,imp=imp.E,model=Model.B.mis)
```

Figure 6 shows the resulting histogram with D for the predicted datasets. Each red line in the histogram shows a score D for an imputed new dataset. Most D values for imputed new data occur in the second half of D-scores for the predicted data. Associated to each D for the imputed datasets is a prior predictive p-value. The distribution of prior predictive p-values is shown in Figure 7.

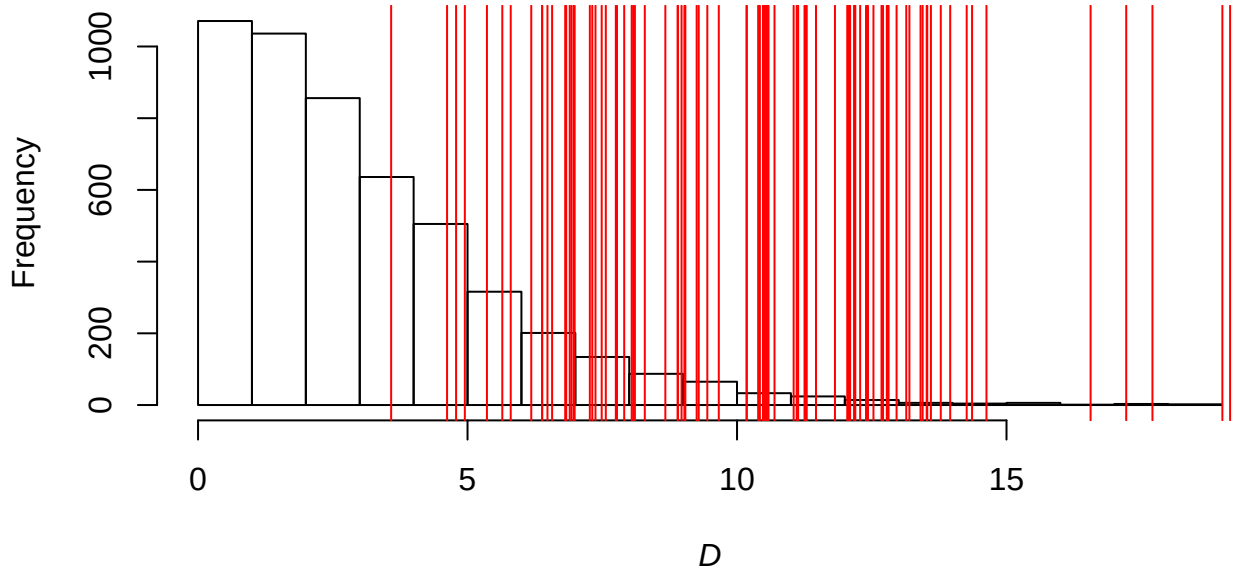

Figure 6. Histogram of D for the predicted data with scores for the imputed new datasets in red.

The computed p-values over the imputations vary from 0.00 to 0.47. 57.0 percent of the prior predictive p-values was smaller than .05. Significant p-values indicate that the new study shows extreme misfit considering H_0 relative to the predicted data given the original study. The proportion of non-significant p-values must be interpreted in light of the fact that the new study did not guarantee high statistical power, as the sample was not particularly large and even smaller than in the original study.

All in all, the outcomes make us skeptical about the replication of the most important findings of Bakker et al. (2013) as captured in H_0.

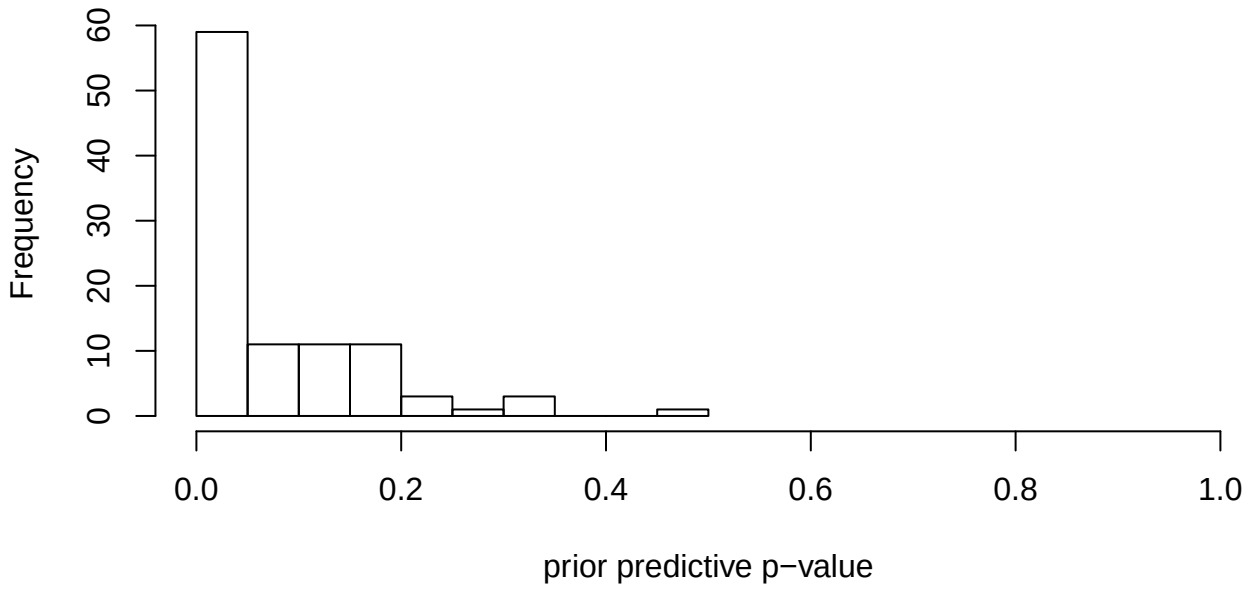

Figure 7. Histogram of prior predictive p-values for the imputed datasets.

Dicussion and Conclusion

The `Replication` R-package enables researchers to test replication for models ranging from simple regressions and ANOVAs up to multilevel structural equation models with missing data. For simple situations, researchers only need to define the model and the replication hypothesis, and run two functions from the `Replication` package. In more complex situations, the `Replication` package provides additional `lavaan` modeling options, and it can handle imputed data.

A future direction for replication testing with the prior predictive p-value would be to delve deeper into the issue of missing data in $\boldsymbol{y}_r$. For example, it would be preferable if we could arrive at one (pooled) prior predictive p-value. Furthermore, power and required sample sizes for the prior predictive p-value can be calculated once we can define the alternative (non-replication) population (Zondervan-Zwijnenburg et al., 2019). In structural equation models, simply setting all H_0 parameters at 0 in H_a however, may result in non-positive definite variance-covariance matrices and non-convergence.

With respect to the replication hypothesis, it may occur that multiple definitions of H_0

seem defensible. In that case researchers could evaluate multiple replication hypotheses and show to what degree the new study fails to replicate the findings in the original study. If researchers were to evaluate multiple specifications for H_0, it is essential that they determine the operationalizations of H_0 before running the analyses, and that they report all investigated H_0 and associated results. Changing H_0 or reporting a selection of the results would be unethical and undermine the goal of replication studies altogether. Thus, we urge scientists to be open about their decisions and investigations.

The current paper demonstrated the use of the `Replication` package with two examples. The Supplementary Material provides data and R-scripts to follow each step in this paper. This facilitates readers who want to test the replication of study claims, including and beyond effect sizes, for any structural equation model.

References

Achterberg, M., Duijvenvoorde, A. C. K. van, Van der Meulen, M., Bakermans-Kranenburg, M. J., & Crone, E. A. (2018). Heritability of aggression following social evaluation in middle childhood: An fMRI study. *Human Brain Mapping*, (1-14). https://doi.org/10.1002/hbm.24043

Achterberg, M., Duijvenvoorde, A. C. K. van, Van der Meulen, M., Euser, S., Bakermans-Kranenburg, M. J., & Crone, E. A. (2017). The neural and behavioral correlates of social evaluation in childhood. *Developmental Cognitive Neuroscience*, *24*, 107–117. https://doi.org/10.1016/j.dcn.2017.02.007

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods*, *21*(1), 1–12. https://doi.org/10.1037/met0000051

Bakker, A., Van der Heijden, P. G., Van Son, M. J., & Van Loey, N. E. (2013). Course of traumatic stress reactions in couples after a burn event to their young child. *Health Psychology*, *32*(10), 1076–1083. https://doi.org/10.1037/a0033983

Box, G. E. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, *143*(4), 383–430. https://doi.org/10.2307/2982063

Carleton, R. N., Gosselin, P., & Asmundson, G. J. (2010). The intolerance of uncertainty index: Replication and extension with an english sample. *Psychological Assessment*, *22*(2), 396. https://doi.org/10.1037/a0019230

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Methods*, *22*(2), 240–261. https://doi.org/10.1037/met0000065

Egberts, M. R., Schoot, R. van de, Geenen, R., & Van Loey, N. E. (2017). Parents' posttraumatic stress after burns in their school-aged child: A prospective study. *Health Psychology*, *36*(5), 419–428. https://doi.org/10.1037/hea0000448

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC. https://doi.org/10.2307/2965436

Harms, C. (2018). A bayes factor for replications of anova results. *The American Statistician*. https://doi.org/10.1080/00031305.2018.1518787

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press. https://doi.org/10.1201/b11158

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2018). Replication bayes factors from evidence updating. *Behavior Research Methods*, 1–11. https://doi.org/10.3758/s13428-018-1092-x

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

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*, *9*, 599–620. https://doi.org/10.1207/S15328007SEM0904/_8

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

R Core Team. (2017). *R: A language and environment for statistical computing* (3.4.0 ed.). Vienna, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004). To Bayes or not to Bayes, from whether

to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 424–451. https://doi.org/10.1207/s15328007sem1103/_7

Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: Order, inequality, and shape constraints* (Vol. 912). John Wiley & Sons. https://doi.org/10.1002/9781118165614

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science*, *9*(3), 305–318. https://doi.org/10.1177/1745691614528518

Stokes, J. M., Pogge, D. L., & Zaccario, M. (2013). Response character styles in adolescents: A replication of convergent validity between the mmpi–a and the rorschach. *Journal of Personality Assessment*, *95*(2), 159–173. https://doi.org/10.1080/00223891.2012.730084

Turlach, B. A., & Weingessel, A. (2013). *Quadprog: Functions to solve quadratic programming problems* (R package version 1.5-5). Retrieved from https://CRAN.R-project.org/package=quadprog

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2013). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development*, *85*(3), 842–860. https://doi.org/10.1111/cdev.12169

Zondervan-Zwijnenburg, M., Van de Schoot, R., & Hoijtink, H. (2019). Testing ANOVA replications by means of the prior predictive p-value. *PsyArXiv*. https://doi.org/10.31234/osf.io/6myqh

Appendix A

```
model.A <- '
#latent growth model
i  =~ 1*positive + 1*neutral + 1*negative #latent factor intercept
s1 =~ 0*positive + 1*neutral + 1*negative #latent factor slope 1
s2 =~ 0*positive + 0*neutral + 1*negative #latent factor slope 2

i  ~ 1        #baseline / first mean
s1 ~ (s1)*1   #dif 12. mean 2 = i+s1
s2 ~ (s2)*1   #dif 23. mean 3 = i+s1+s2

#residual variances repeated measures
positive ~~ (rt1)*positive
neutral  ~~ (rt2)*neutral
negative ~~ (rt3)*negative

#item means @0
positive ~0*1
neutral  ~0*1
negative ~0*1

#(co)variances latent factors @0
i  ~~ 0*i          #fixed intercept factor
s1 ~~ 0*s1         #fixed s1 factor
s2 ~~ 0*s2         #fixed s2 factor
i  ~~ 0*s1 + 0*s2  #no covariance i  & s1, i & s2
```

```
s1 ~~ 0*s2          #no covariance s1 & s2


#pooled standard deviations
s12 := sqrt((rt1+rt2)/2)
s23 := sqrt((rt2+rt3)/2)
s13 := sqrt((rt1+rt3)/2)


#Cohens d effect sizes
d12 := s1      /s12
d23 := s2      /s23
d13 := (s1+s2) /s13
'
```

In the model syntax above, the operator `=~` defines a latent factor, the operator `~1` indicates a regression on one, which is used for means and intercepts. The operator `~~` is used for (co)variances between the variable at the left hand side and the right hand side. Before a `*`, labels and fixed values can be inserted. The pooled standard deviations and the effect sizes are included in the model as defined parameters with the operator `:=`, which means that they are not estimated, but they are derived from other estimated parameters.

Appendix B

The following prior has been used in the analysis of Achterberg et al. (2017) for the latent factors $\boldsymbol{\alpha}$.

$$\alpha \sim N(0.00, 0.01),$$

which denotes a normal distribution with a mean of 0 and a precision (i.e., the inverse of the variance) of 0.01. A visualization of this default `blavaan` prior is depicted in Figure 8.

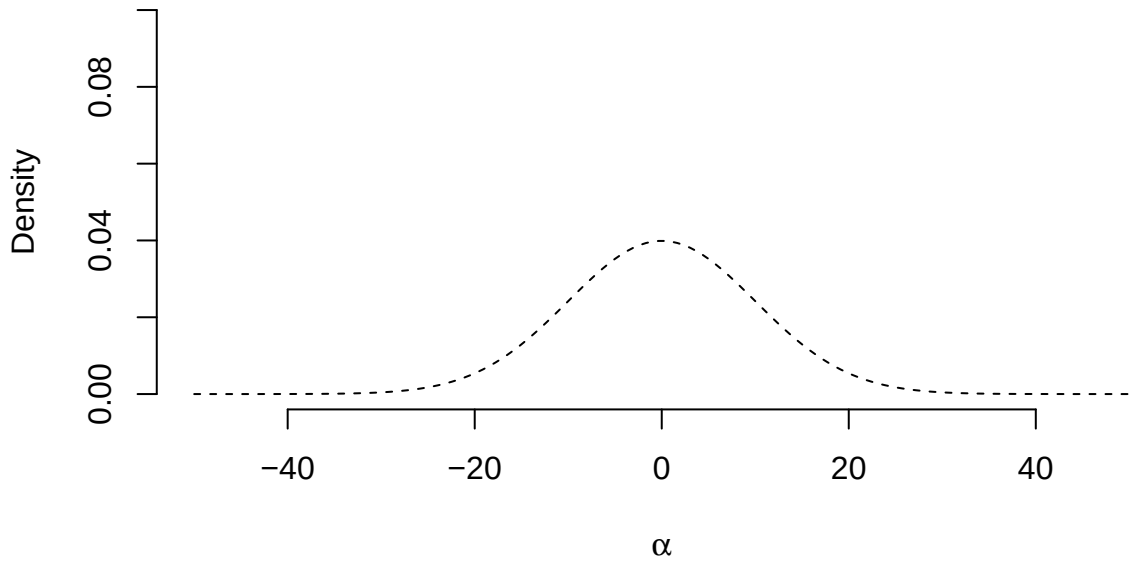

Figure 8. Prior distribution for $\boldsymbol{\alpha}$.

The following prior has been used in the analysis of Achterberg et al. (2017) for the residual variances $\boldsymbol{\Theta}$.

$$\Theta \sim \Gamma(1.00, 0.50),$$

which denotes a gamma distribution with a shape parameter of 1.00, and a rate (i.e., the inverse of the scale) of 0.50. A visualization of this default `blavaan` prior is depicted in Figure 9.

The default priors for all model parameters in `blavaan` can be consulted with:

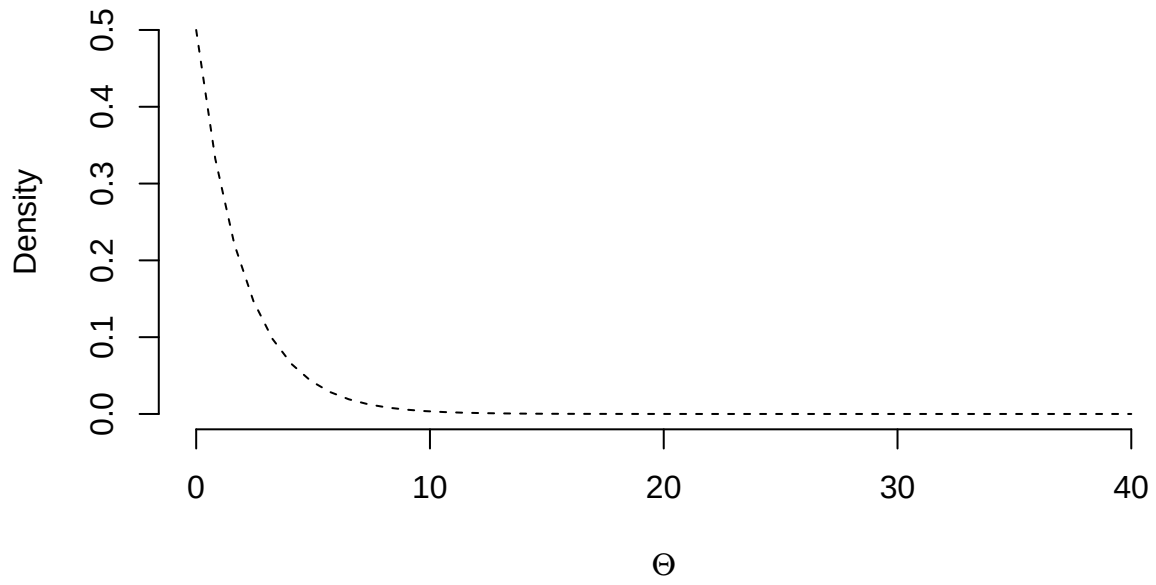

Figure 9. Prior distribution for $\mathbf{\Theta}$.

```
dpriors()
```

```
##                nu             alpha            lambda              beta
## "dnorm(0,1e-3)" "dnorm(0,1e-2)" "dnorm(0,1e-2)" "dnorm(0,1e-2)"
##            itheta              ipsi               rho             ibpsi
##  "dgamma(1,.5)"  "dgamma(1,.5)"    "dbeta(1,1)" "dwish(iden,3)"
##               tau             delta
##   "dnorm(0,.1)"  "dgamma(1,.5)"
```

These are also the priors used to evaluate the replication of Bakker et al. (2013).