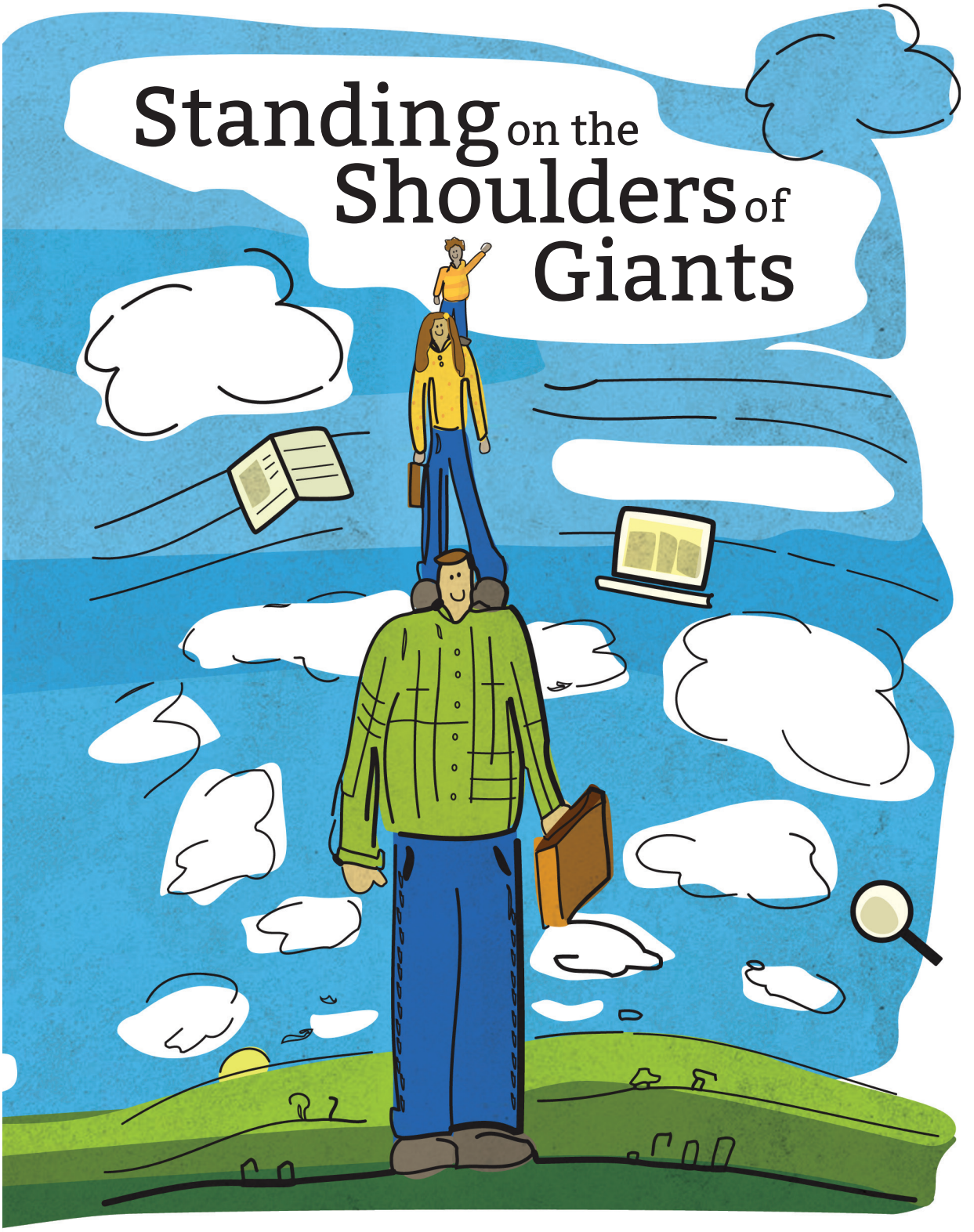
Standing on the Shoulders of Giants



Formalizing and Evaluating Prior Knowledge

By Mariëlle Zondervan-Zwijnenburg

STANDING ON THE SHOULDERS OF GIANTS: FORMALIZING AND EVALUATING PRIOR KNOWLEDGE

STAAN OP DE SCHOUDERS VAN REUZEN:
FORMALISEREN EN EVALUEREN VAN VOORKENNIS
(met een samenvatting in het Nederlands)

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op vrijdag 4 oktober 2019 des middags te 4.15 uur

door

Maria Adriana Joëlle Zondervan-Zwijnenburg

geboren op 24 december 1989
te Lopik

Promotoren: Prof. dr. H.J.A. Hoijtink
Prof. dr. A.G.J. van de Schoot

This research was financially supported by The Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003).

Beoordelingscommissie:

Dr. G. Altoè
Prof. dr. I.G. Klugkist
Prof. dr. M.A.L.M. van Assen
Prof. dr. W.A.M. Vollebergh
Prof. dr. L.A. van der Ark

Zondervan-Zwijnenburg, Mariëlle
Standing on the Shoulders of Giants: Formalizing and Evaluating Prior Knowledge
Proefschrift Universiteit Utrecht, Utrecht. - Met lit. opg. - Met samenvatting in het Nederlands.
ISBN 978-94-6332-545-5
Druk: GVO drukkers & vormgevers B.V. | Ponsen & Looijen
Cover design by Jessica Muñiz-Winter @colorhive

Contents

1

Introduction

> The dwarf sees farther than the giant,
> when he has the giant's shoulder to mount on.
>
> *Samuel Taylor Coleridge*
> (1828)

The citation above has been used to express that we can learn more about the truth by building on previous discoveries. Building on discoveries and evaluating the worth of previous discoveries is what this dissertation is about. Two manners to include previous discoveries in statistical analyses are informative priors and informative hypotheses, which will be shortly explained in section 1.1 and 1.2 respectively.

1.1 Informative Priors

Informative priors can be used within Bayesian statistics. Bayesian statistics are named after reverend Thomas Bayes (1701-1761) and were further developed by Pierre-Simon Laplace (1749-1829). In Bayesian statistics, the probability of hypotheses or estimates in parameter vectors are based on: (1) prior knowledge, and (2) the current data. If we denote de hypothesis by H and the data by D, then Bayes' theorem is given by:

$$P(H|D) = \frac{P(H)P(D|H)}{P(D)}, \tag{1.1}$$

where $P(H|D)$ is the probability of the hypothesis given the data, $P(H)$ is the probability of the hypothesis (i.e., the prior), $P(D|H)$ is the likelihood of the data given the hypothesis, and $P(D)$ is the probability of the data. In case of parameter estimation, H is replaced by the parameter vector $\boldsymbol{\theta}$.

Nowadays, we can apply Bayesian statistics on complex models with many parameters such as means, variances, correlations, and regression parameters. The prior distribution represents the prior information that we have, irrespective of the data at

This chapter is written by Mariëlle Zondervan-Zwijnenburg

hand, about the hypothesis of interest or about the parameters in the model. Where do we get this prior information? To determine the content of the prior, we can make use of previous discoveries presented in earlier research (e.g., Gelman et al., 2013, Chapter 5), or knowledge that we elicit from experts (e.g., O'Hagan et al., 2006).

Consider the following hypothetical example in which we want to estimate the mean of IQ in the Netherlands: μ_{IQ}. From previous research conducted in various European countries, we know that the average IQ is 100. To make use of this knowledge, we can compose a prior distribution. As a prior, we specify a normal distribution with a mean of 100 and some standard deviation, for example, 10. That is:

$$\mu_{\text{IQ}} \sim N(100, 10). \tag{1.2}$$

This prior expresses that the expected value for μ_{IQ} is 100, but we are not completely sure that 100 will be the exact mean in our current population, therefore, we set the standard deviation at 10 instead of a value close to 0. If we would have more doubt about μ_{IQ} being (close to) 100, we would increase the standard deviation of the prior for μ_{IQ}. A visualization of the prior distribution in Equation 1.2 is provided in Figure 1.1. Chapter 3 and 4 respectively demonstrate how prior knowledge can be obtained from previous research and experts, and subsequently used in a Bayesian analysis.

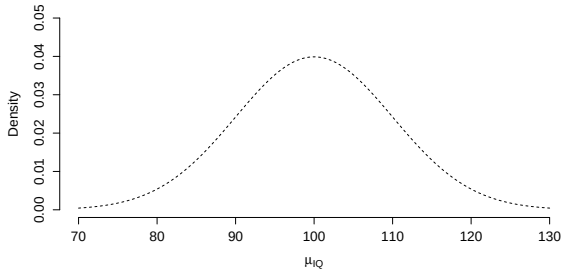


Fig. 1.1: Normal prior distribution for μ_{IQ} with prior mean = 100 and prior standard deviation = 10.

How can we evaluate whether a prior specification is appropriate? In Chapter 2, the impact of the prior distribution is evaluated in a simulation study. In Chapter 3, multiple analyses with varying priors are conducted to evaluate the impact of the prior distribution. In Chapter 4, the prior distribution, the likelihood of the data, and the posterior solution are visualized to clarify the impact of the prior distribution. In Chapter 5, 6 and 7, the prior distribution is based on a study for which a replication effort has been made. The prior predictive check Box (1980) is used here to evaluate whether the new study fails to replicate relevant features of the original study. Those features are captured in an informative hypothesis, which is the topic of the next section.

1.2 Informative Hypotheses

Another way to incorporate prior knowledge are informative hypotheses (Silvapulle and Sen, 2005; Hoijtink, 2012). Typically, researchers evaluate a classical null hypothesis H_0 which captures the situation that "nothing is going on", for example, H_0: $\mu_1 = 0, \mu_2 = 0$, or $H_0 : \mu_1 = \mu_2$. In contrast, informative hypotheses H_i express researchers' expectations about the parameters. These expectations can be expressed through range constraints, order constraints, and equality constraints. Consider again a hypothetical example in which we evaluate the mean IQ of pupils in regular education $\mu_{\text{IQ-r}}$ and special education $\mu_{\text{IQ-s}}$. With range constraints we imply that the parameters cannot take on any value, but only values within a certain range, that is:

$$H_i : 85 < \mu_{\text{IQ-r}} < 115, 60 < \mu_{\text{IQ-s}} < 90.$$

With an order constraint, we order the parameters in the model, that is:

$$H_i : \mu_{\text{IQ-r}} > \mu_{\text{IQ-s}}.$$

With an equality constraint, we can specify equality between parameters, or constrain the value of parameters, that is:

$$H_i : \mu_{\text{IQ-r}} = \mu_{\text{IQ-s}}, \text{or } H_i : \mu_{\text{IQ-r}} = 100, \mu_{\text{IQ-s}} = 75.$$

For informative hypotheses it also holds that their content can be based on previous discoveries presented in earlier research or knowledge held by experts. In Chapter 5, 6, and 7 the conclusions of an earlier study determine the content of the informative hypothesis. In Chapter 8, informative hypotheses are based on exploratory discoveries from multiple cohort studies.

To test an informative hypothesis, we can compute (1) a frequentist p-value (see Silvapulle and Sen, 2005), (2) an information criterion comparing different informative hypotheses (see Kuiper and Hoijtink, 2013), or (3) a Bayes factor comparing different hypotheses (see Gu et al., 2018). In Chapters 5, 6 and 7, the informative hypothesis leads to test statistics for which we calculate a prior predictive p-value (Box, 1980). In Chapter 8, we use the Bayes factor to evaluate the relative amount of evidence for competing informative hypotheses.

1.3 Aim and Outline

The aim of this dissertation is to demonstrate how prior knowledge can be formalized and evaluated. Part I concentrates on acquiring prior knowledge for Bayesian analyses. Part II introduces testing replication by means of the prior predictive p-value, and in Part III, exploratory analyses lead to informative hypotheses that are evaluated with Bayes factors. The dissertation ends with a Discussion chapter in which I highlight the contribution of the different chapters, and discuss what future research may aim for.

Part I

Acquiring Prior Knowledge

2nd

2

Pushing the Limits: The performance of ML and Bayesian estimation with small and unbalanced samples in a latent growth model

Summary. Longitudinal developmental research is often focused on patterns of change or growth across different (sub)groups of individuals. Particular to some research contexts, developmental inquiries may involve one or more (sub)groups that are small in nature and therefore difficult to properly capture through statistical analysis. The current study explores the lower-bound limits of subsample sizes in a multiple group latent growth modeling by means of a simulation study. We particularly focus on how the maximum likelihood (ML) and Bayesian estimation approaches differ when (sub)sample sizes are small. The results show that Bayesian estimation resolves computational issues that occur with ML estimation, and that the addition of prior information can be the key to detect a difference between groups when sample and effect sizes are expected to be limited. The acquisition of prior information with respect to the smaller group is especially influential in this context.

Many researchers in the social and behavioral sciences use latent growth modeling (LGM) to study development of individuals over time (e.g., Little, 2013). Within LGM it is also possible to compare growth and the impact of variables on growth between different groups of individuals, for example, between a focal (i.e., small) group and a reference group. Researchers with this objective, however, often encounter two difficulties. In particular, the comparisons they want to make are between groups: (1) that have relatively different sample sizes, or (2) of which at least one is considered to be very small according to common guidelines for implementing the statistical model.

This chapter is published as Zondervan-Zwijnenburg, M.A.J., Depaoli, S., Peeters, M., & Van de Schoot, R. (2018). Pushing the Limits: The performance of ML and Bayesian estimation with small and unbalanced samples in a latent growth model. *Methodology, 15*, 31-43. doi: 10.1027/1614-2241/a000162.

Author contributions: MZ and RS designed the study. MZ, RS, and SD contributed to the design of the simulation study. MP collected the data. MZ wrote the paper under supervision of RS. Additional feedback was provided by SD and MP.

From the literature, we know that with traditional maximum likelihood (ML) estimation, the consequences of small sample sizes can include: biased point estimates (Boomsma and Hoogland, 2001; Depaoli, 2013; Lee and Song, 2004; Lüdtke et al., 2011; Meuleman and Billiet, 2009; Van de Schoot et al., 2015), inadmissible estimates (Boomsma and Hoogland, 2001; Can et al., 2015; Hox and Maas, 2001; Meuleman and Billiet, 2009; Tolvanen, 2000), convergence issues (Boomsma and Hoogland, 2001; Hochweber and Hartig, 2017; Hox et al., 2014; Lüdtke et al., 2011), and inflated Type-I error rates (Boomsma and Hoogland, 2001; Hox and Maas, 2001; Hox et al., 2014; Lee and Song, 2004; Meuleman and Billiet, 2009).

There is, however, little known about the consequences of unbalanced samples (i.e., where sample sizes vary drastically across the subgroups being examined, e.g., 10 participants in the focal group vs. 500 in the reference group), especially when latent growth models are being implemented. We only know that unbalanced samples in LGM often result in low statistical power (Muthén and Curran, 1997), but its specific effect on coverage, biased point estimates, and estimation problems has not been thoroughly examined in the literature. Altogether, these issues may deter researchers from comparing the development of focal and reference groups in latent growth models.

Bayesian estimation is an alternative estimation method gaining in popularity (Kruschke, 2011; Van de Schoot et al., 2017). In Bayesian statistics, prior information is combined with the data in the analysis, resulting in a posterior distribution. The posterior distribution reflects probable parameter values given the prior information and the data. From the posterior distribution, a measure of central tendency (i.e., the mean, median, or mode) is usually taken as a point estimate for the parameter of interest. Additionally, a 95% (credible) interval can be derived from the posterior distribution containing the most probable values for the parameter given the data. The frequentist 95% confidence interval, in contrast, will contain the true population value in 95% of the intervals over a long run of trials. To readers interested in a gentle introduction into Bayesian statistics for social scientists, we recommend Kruschke (2014), and Van de Schoot et al. (2013).

In the current paper, we conduct a simulation study to evaluate the performance of maximum likelihood estimation and Bayesian estimation for latent growth models with small and unbalanced samples. The goal of the simulation is to highlight best practice when dealing with subgroup sizes that are quite different from one another.

2.1 Background on Sample Size Limits in LGM with ML and Bayesian Estimation

Muthén and Curran (1997) investigated the effect of unbalanced sample sizes in experimental designs on statistical power in LGM with sample size ratios varying from 1:1 (balanced) to 1:10. In general, Muthén and Curran (1997) found that the more extreme the sample size ratios were, the lower the statistical power to detect a difference between groups with ML estimation. When the ratio was more extreme than 1:5, even samples with 1,000 participants in total showed less than desirable power

(<.80) to detect a small effect (Cohen's $d = .20$). Due to their focus on experimental designs, Muthén and Curran (1997) do not cover very small sample sizes, extreme sample size ratios, or the inclusion of covariates to limit the impact of confounders. No literature was found that covered aspects other than power under unbalanced sample sizes in LGM.

With respect to estimation in relation to total sample size for one group, estimates from ML estimation with a sample size as low as 50 do not substantially deviate from the population value (i.e., relatively unbiased) for means and factor loadings in LGM and related multilevel models (Hox and Maas, 2001; Maas and Hox, 2005; McNeish, 2016a,b; Meuleman and Billiet, 2009; Tolvanen, 2000). Statistical power, however, is generally insufficient with samples smaller than 100 for the types of effect sizes commonly seen in empirical studies, and convergence issues also arise (Boomsma and Hoogland, 2001; Hochweber and Hartig, 2017; Hox and Maas, 2001; Lüdtke et al., 2011; Maas and Hox, 2005; Meuleman and Billiet, 2009; Tolvanen, 2000). Bayesian estimation does not have the same issues with small samples as ML estimation for two reasons. First, in Bayesian estimation, the results are determined by more than the data: prior information is also included by means of prior distributions. Prior distributions can be based on information that a researcher has about parameters in the model a priori. When no information is available, so-called uninformative distributions can be adopted, which typically specify a very wide range of values for the parameter as probable. The more prior mass surrounding the population value, the better the model estimate will represent this value. Consequently, the non-null detection rate is higher, and inference errors are less likely to occur (Lee and Song, 2004; Depaoli, 2013; Van de Schoot et al., 2015).

The second reason Bayesian estimation does not have the same issues with small samples is that Bayesian estimation does not rely on asymptotic assumptions about sampling distributions akin to ML estimation (Asparouhov and Muthén, 2010). Depaoli (2013) shows in a growth mixture model that the use of uninformative priors as compared to ML estimation results in fewer problematically biased parameter estimates (i.e., bias $\geq 10\%$). When Bayesian estimation is used with an uninformative prior, a sample size of 20 already results in accurate estimates in a multilevel model (Hox et al., 2012). In addition, the coverage of the population value was better with Bayesian estimation, a result confirmed by Van de Schoot et al. (2015) for repeated-measures analyses.

2.1.1 The Current Investigation

In order to ensure conditions were applicable to real data situations, the simulation study is inspired by an empirical dataset on the development rate of working memory in young heavy cannabis users versus their non-using peers. The data originate from

Statistical power is a frequentist term that involves the null hypothesis. Since the null hypothesis does not exist in Bayesian statistics, we refer to the non-null detection rate instead.

268 young adolescents enrolled in special education due to behavioral problems (Peeters et al., 2014). To improve on the notion of causality, the development of both groups was corrected (by means of a time-invariant covariate) for the impact of quantity and frequency of alcohol use at the start of the study, as recommended by Jacobus et al. (2009). We set up the simulation this way in order to compare and establish sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation when one of the groups has a sample size below 50.

By means of the simulation, we compare the sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation. Regarding Bayesian estimation, the balance between sample size requirements and the required specificity of prior information is investigated as well. Additionally, we explore how the results are affected when a substantial amount of prior information can be found for the reference group but not for the focal group. It can be expected that prior information with respect to a focal group is harder to obtain.

2.2 Method

To compare the performance of ML estimation and Bayesian estimation in the evaluation of small and unbalanced samples in a latent growth model, we conducted a Monte Carlo simulation study in Mplus version 7.11 (Muthén and Muthén, 2012) directed by the R-package MplusAutomation (Hallquist, 2013) in R 3.0.1 (R Core Team, 2013). To promote transparency and replicability, analyses syntax files and all input and output is available at `osf.io/gjzu8`. In this section, we elaborate on the model of interest, the main characteristics of the simulation study, and the evaluation criteria.

2.2.1 The Latent Growth Model

Figure 2.1 displays the latent growth model as applied in this study. The model has four observed variables ($y_1^g - y_4^g$) representing repeated measures of the same construct. In the empirical data, this construct is performance on a working memory task expressed in percentages. The repeated measures are the indicators for the intercept, linear slope, and quadratic slope latent variables. The linear growth factor in this model represents the growth rate at one time point (typically the first time point). The model has one covariate representing an observed time-invariant predictor, which is a measure of alcohol use quantity and frequency at the start of the study in the empirical data. As a result, the latent time variables technically have intercepts instead of means. However, to avoid confusion between the intercept growth factor and the intercepts of the latent growth factors, the latter will be referred to as being "means" throughout the paper.

In order to assess the growth rate difference between groups, a new parameter (denoted by $\Delta\alpha$) was constructed by subtracting the linear slope mean of group 2 (i.e., the focal group) from that of group 1 (i.e., the reference group).

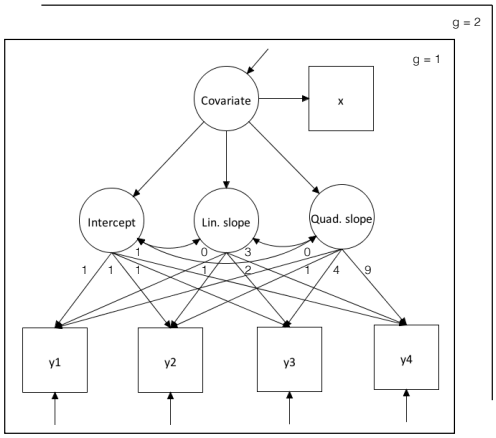


Fig. 2.1: Multiple group latent growth model with one covariate and groups indicated by g. y_1^g, y_2^g, y_3^g, y_4^g represent four assessments of a developing construct with residual error variances. x^g is a time-invariant predictor of growth that represents the latent variable Covariate^g without measurement error. The regressions of the latent growth factors Intercept^g, Lin. slope^g, and Quad. slope on the Covariate^g are equal over groups.

2.3 Simulation Study Design

The population parameters originated from multiple group latent growth analyses (see `osf.io/ttybt`) on empirical data. The difference between the linear slope factors, $\Delta\alpha$, was set at 1.60, while the disturbance of the linear slope factors was 64.00 in order to represent a small effect size ($\frac{1.60}{\sqrt{64.00}} = .20$ Cohen's d; (Cohen, 1988)), which is considered a realistic effect size for this parameter (see, for instance, Jacobus et al., 2009).

For this population, we varied the sample sizes in the reference group, the sample sizes in the focal group, and the estimation settings. The sample sizes for the reference group were $\in \{50, 100, 200, 500, 1,000, 2.000, 5.000, 10.000\}$, which represents a wide range of sample sizes commonly specified in the empirical and methodological literature. The sample sizes for the focal group were 5, 10, 25, and 50. Consequently, the sample size ratios ranged from 1:1 to 1:2,000. The estimation methods were ML estimation and Bayesian estimation.

ML estimation was applied with standard errors robust to non-normality and non-independence of observations (MLR), which suits analyses with repeated measures. Mplus uses accelerated expectation maximization (EMA) to obtain the ML estimates (Muthén and Muthén, 2012). Syntax for the analyses is provided at `osf.io/gjzu8`. The ML output shows one extra parameter compared to the exact same Bayesian specification. This "knownclass" parameter, however, is not estimated. Therefore, we consider the models to be exactly equal.

Bayesian estimation was implemented with seven different prior distribution settings for the means of the latent growth factors. Normally distributed informative priors were specified for the latent growth factor means, because it was considered most likely that researchers would have knowledge about these parameters before analyzing their data. Theoretically, however, prior information can be found for all parameters. The more appropriate the information being included in the prior is, the more accurate the parameter estimates will be. All user-specified priors were normally distributed with mean μ_0 and variance σ_0^2. The population values of the growth factor means were used as prior means to understand the upper-bound performance of Bayesian methods under these modeling circumstances. The prior variances σ_0^2 ranged from 0.1 (i.e., very informative) to 10^{10} (i.e., uninformative). Specifically, $\sigma_0^2 \in \{0.1, 0.3, 0.5, 1.0, 2.0, 5.0, 10^{10}\}$. Default priors were used for the other parameters in the model. Specifically:

- A normal distribution with a mean of 0 and variance of 10^{10} for the mean of the covariate and the regression coefficients,
- An improper inverse gamma with the shape parameter set at -1, and the scale at 0 for the variance of the covariate and the residuals of the observed variables,
- An improper inverse Wishart with 0 forming the scale matrix, and -4 degrees of freedom for the covariances and residual variances of the growth factors.

Furthermore, 22 Markov chains were used for the Bayesian analyses to capture the impact of many different starting values. Note, however, that 22 chains may be excessive in other modeling contexts due to the length of time it would take to obtain convergence. We were able to have the large number due to the computational capacity that was available to us. It is important to note that methods and results described here using these 22 chains are generalizable to situations requiring fewer chains. In order to assess convergence, it is recommended that at least two chains are used (Gelman and Rubin, 1992). The minimum number of iterations was set at 5,000, and the maximum at 100,000. The first half of the chain was discarded as burn-in, and the second half was used to construct the posterior (Muthén and Muthén, 2012).

Convergence was imposed by means of the Gelman-Rubin potential scale reduction factor (PSRF; Gelman and Rubin, 1992). When the PSR was less than 0.05 points away from 1 for all parameters in the second half of the iterations, the model was considered to be converged. Subsequently, the first half of the iterations was discarded as a burn-in phase (Muthén and Muthén, 2012). Syntax for the analyses is provided at `osf.io/gjzu8`. Altogether, the number of cells in the simulation study was 4 (focal group sample sizes) $\times$ 8 (reference group sample sizes) $\times$ 8 (estimation settings: 1 $\times$ ML + 7 $\times$ Bayes with varying σ_0^2) = 256.

The simulation was extended with additional Bayesian analyses to investigate what would happen if a substantial amount of prior information (specified as having a variance hyperparameter of $\sigma_0^2 = 0.1$, indicating a great deal of precision in the prior)

That is, 73.05, 71.54, 8.13, 6.53, and -2.16 for $\text{Intercept}_{\text{non-users}}$, $\text{Intercept}_{\text{users}}$, Lin. $\text{slope}_{\text{non-users}}$, Lin. $\text{slope}_{\text{users}}$, and Quad. slope, respectively

could only be obtained for the reference group, but not for the focal group (with a variance hyperparameter of $\sigma_0^2 = 10.0$, indicating less precision in the normal prior). In the focal group σ_0^2 was set at 10.0 instead of 10^{10} (the Mplus default) because, even when prior information is hard to find, researchers and experts are generally able to estimate its value to some extent. We investigated the effects of these conditions for the largest (i.e., best performing) focal group ($n = 50$). The sample size of the reference group was again manipulated for this additional condition examined. Input for this analysis is located at `osf.io/xm3v5`

2.4 Evaluation

Since the main interest in multiple group LGM is to compare development between groups, the growth rate difference parameter $\Delta\alpha$ was the parameter of interest in the simulation study. For the Bayesian cells in the design, the median of the posterior distribution was interpreted as the point estimate. Credible intervals were obtained by the equal tail method, having tails on both sides that each contain 2.5% of the posterior distribution (Muthén and Muthén, 2012).

The difference parameter $\Delta\alpha$ was evaluated in terms of proportion of bias, coverage, statistical power or non-null detection rates, and estimation problems. The proportional bias was calculated by dividing the average bias over the analyzed datasets by the value of the population estimate. A proportional bias lower than .10 was considered acceptable (Muthén and Muthén, 2002). Coverage is the rate of 95% confidence intervals (frequentist statistics, e.g., ML estimation) or credible intervals (Bayesian statistics) that covers the population parameter estimate. For a 95% confidence or credible interval, coverage should be around the advocated 95%. In the current study, a minimum level of .90 was considered acceptable. Statistical power and non-null detection rates were calculated as the percentage of replications in which the 95% interval for $\Delta\alpha$ did not include zero. The acceptable minimum level of statistical power or the non-null detection rate was considered to be .80 (Muthén and Muthén, 2002). The last criterion concerned estimation problems. Estimation problems arise when the following occur: (1) negative variances, (2) correlations larger than one, (3) linear dependencies among more than two latent variables are estimated, or (4) when the model does not converge. When using ML estimation, Mplus notifies the user when one of these problems occurred. The proportion of datasets for which Mplus produced warnings in this respect was used as an evaluation of estimation problems. Bayesian estimation cannot result in illegitimate estimates with the prior distributions used in this study. Non-convergence, however, can occur, and can be detected by warnings and/or by visual inspection of the trace plots. Therefore, for every cell in the simulation design, two sets of trace plots were randomly selected and inspected for convergence.

2.5 Results

2.5.1 Maximum Likelihood Estimation

Figure 2.2 shows the ML results in terms of proportion of warnings, coverage, statistical power, and proportional bias for the four focal group sample sizes separately. As can be seen, the proportion of bias was adequate for all combinations of sample sizes, except for a focal group sample size of 5 combined with a reference sample of 100 (Figure 2.2a). Coverage was in general lower than .95, but always sufficient when the focal sample contained at least 25 participants (Figure 2.2c, 2.2d). With sample sizes in the focal group of 5 and 10, reference group sample sizes at both extreme ends did not cover the population value often enough in the 95% confidence intervals (coverage $< .90$), even though the average relative bias over datasets was acceptable (Figure 2.2a, 2.2b). Truly worrisome, however, were the statistical power and the proportion of warnings. Even with 10,000 participants in the reference group, the power to detect a small effect was lower than .50 for all focal groups, while a minimum of .80 is pursued. The proportion of warnings with a reference group sample size of 50 ranged from .73 to .88. These warnings concerned illegitimate estimates, which make the results of the analysis unreliable. Examples of warnings that were obtained for ML models with estimation issues were as follows:

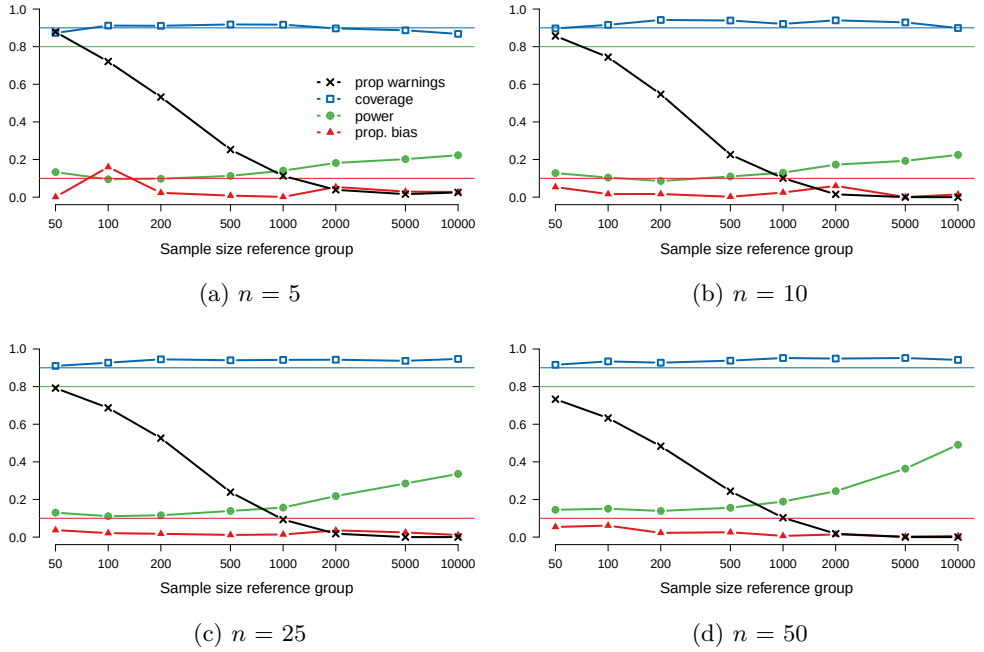
```
THE MODEL ESTIMATION TERMINATED NORMALLY

WARNING:  THE RESIDUAL COVARIANCE MATRIX (THETA) IS NOT POSITIVE DEFINITE.
THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR AN OBSERVED
VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO OBSERVED
VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO OBSERVED VARIABLES.
CHECK THE RESULTS SECTION FOR MORE INFORMATION.

WARNING:  THE LATENT VARIABLE COVARIANCE MATRIX (PSI) IS NOT POSITIVE
DEFINITE.  THIS COULD INDICATE A NEGATIVE VARIANCE/RESIDUAL VARIANCE FOR A
LATENT VARIABLE, A CORRELATION GREATER OR EQUAL TO ONE BETWEEN TWO LATENT
VARIABLES, OR A LINEAR DEPENDENCY AMONG MORE THAN TWO LATENT VARIABLES.
CHECK THE TECH4 OUTPUT FOR MORE INFORMATION.
```

2.5.2 Bayesian Estimation

With Bayesian estimation, bias and coverage were acceptable for every cell of the simulation design. Plots for all cells can be found at `osf.io/s59cz`. In addition, Bayesian estimation showed decent convergence. As a result, the remaining aspect of interest was statistical power. Figure 2.3 shows for all four focal group sample sizes (i.e., $n = 5$, 10, 25, and 50) how many participants are in the reference group and how much prior information is necessary to obtain satisfactory non-null detection rates. With uninformative priors imposed on all parameters (i.e., $\sigma_0^2 = 10^{10}$), non-null detection rates were insufficient, regardless of the sample size in the reference group. The same held when the variances of the priors for the latent growth factor means



(a) $n = 5$ (b) $n = 10$

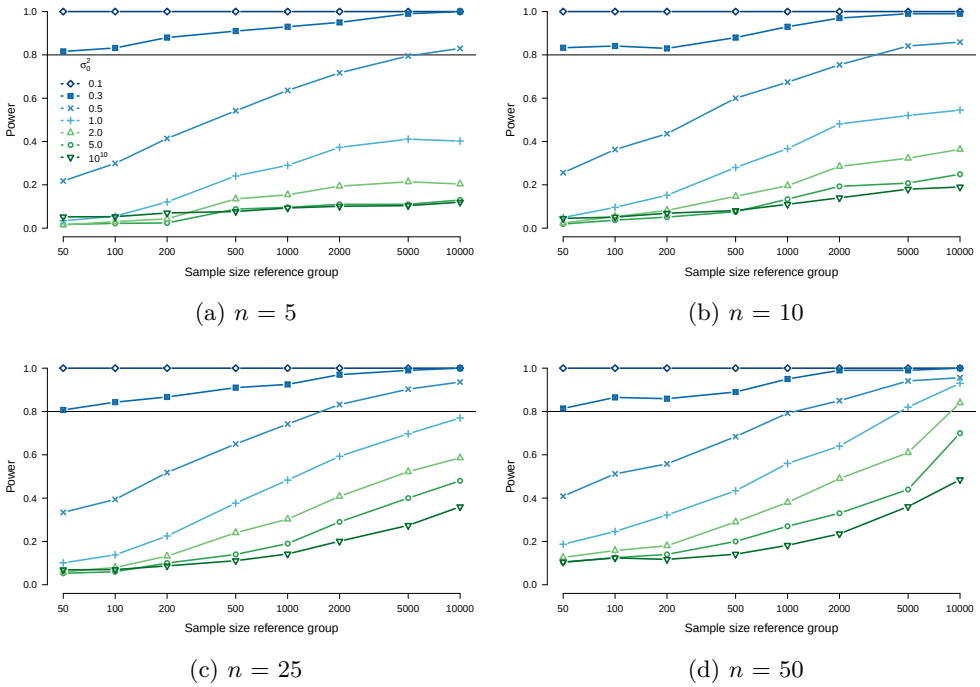
(c) $n = 25$ (d) $n = 50$

Fig. 2.2: Results for ML estimation by focal group sample size. On the x-axis, the size of the reference group increases. From top to bottom, the static horizontal lines represent: (1) the minimum acceptable value for coverage (i.e., .90), (2) the minimum acceptable value for statistical power (i.e., .80), and (3) the maximum acceptable value for proportional bias (i.e., .10).

were decreased to 5.0. An exploration of the non-null detection rate with a focal group of 100 and the prior variance of the latent growth factor means at 5.0 showed an improvement in the non-null detection rate, but still about 10,000 participants in the reference group were needed to acquire a non-nul detection rate close to .80. Prior variances as specific as 0.1, on the other hand, resulted in a non-null detection rate of 1.0 for every cell.

2.5.3 Unbalanced Prior Information

The simulation results presented in the previous section show that an focal group of 50 participants combined with a prior variance is 0.1 can lead to an optimal situation in all respects assessed (Figure 2.3). Figure 2.4 shows that when prior information is scarce for the focal group ($\sigma_0^2 = 10$), power is an issue again. Additional analyses showed that no matter how much the prior variance in the reference group was decreased, a satisfactory non-null detection rate could not be achieved as long as the prior variance



(a) $n = 5$ (b) $n = 10$

(c) $n = 25$ (d) $n = 50$

Fig. 2.3: Non-null detection rate for Bayesian estimation by focal group sample size. On the x-axis, the size of the reference group increases. The y-axis represents the non-null detection rate. The static horizontal line represents the minimum acceptable value for the non-null detection rate (i.e., 0.80). The remaining lines reflect the results for varying σ_0^2.

in the focal group was 10. Due to these clear results, the effect of unbalanced prior information was not further investigated for cells with focal groups smaller than 50.

2.6 Conclusion

The aim of the simulation study was to investigate lower-bound sample size issues in a multigroup LGM context, especially when one group is much smaller than the others. We set up the simulation in this way in order to compare and establish sample size requirements to evaluate a small difference in development between groups for ML and Bayesian estimation when one of the groups has a sample size not larger than 50.

The results showed that ML estimation has issues with statistical power when at least one of the groups is not larger than 50. Moreover, with ML estimation, analyses based on small sample datasets generally cannot be properly interpreted because of nonpositive definite matrices that yield inadmissible estimates.

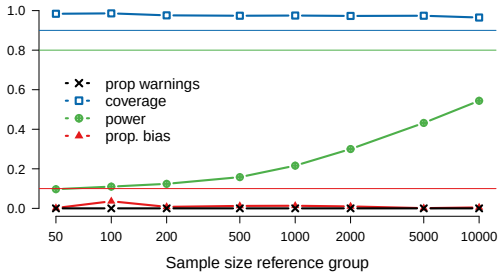


Fig. 2.4: Results for Bayesian estimation with unbalanced prior information. σ_0^2 for latent growth factors in reference group = 0.1. σ_0^2 for latent growth factors in focal group = 10. focal group $n = 50$.

By adopting Bayesian estimation, the issue of non-interpretable output disappears and consequently smaller samples can be analyzed. Bayesian inference with uninformative as well as minimally informative priors, however, has non-null detection rate issues similar to ML estimation. Specifically, even comparison groups with 10,000 participants do not yield satisfactory non-null detection rates for a small effect. To obtain a satisfactory non-null detection rate in the context of limited small and unbalanced sample sizes, Bayesian estimation is necessary in combination with the availability of very specific prior information. This may seem trivial to those who are familiar with the Bayesian concept, but the current simulation study provided additional insight to the effect of prior information by showing the consequences of specific degrees of informativeness.

Note, however, that our use of an empirical model with empirical population values limits the direct applicability of the simulation results to other research situations. The simulation results are only directly indicative for other researchers under specific circumstances. The statistical model needs to be equal (e.g., a latent growth model including a time-invariant covariate, a multiple group confirmatory factor model with a covariate, or a multiple indicators multiple causes model with the groups as a covariate), the expected effect size small, and the growth rate difference needs to be comparable or proportional after taking the impact of the covariate into account. When the growth rate is proportional, the impact of the prior variances is proportional as well. If these circumstances do not hold, the presented simulation results are mainly useful as inspiration for new simulation efforts.

As was shown by the simulation study with unbalanced prior information, highly informative priors are particularly necessary for the focal group. To be able to specify such informative priors, the available prior information must be very specific and convincing. This, however, may be seldom feasible because of the exceptionality of the group. In such a situation, we advise researchers to publish their updated estimates and data nevertheless. Such a publication provides a future researcher on the topic with more prior information, and over time, the amount of prior information can

be sufficient to draw conclusions about the effect under study. Thus, when separate analyses cannot obtain sufficient power to make inferences, cumulative efforts of researchers can overcome the issue.

2.6.1 Cautionary Points Regarding Bayesian Estimation

To avoid misinterpretations of this study, we hereby provide a disclaimer. The goal of Bayesian analyses with informative priors is to make optimal use of all available information. Accordingly, the simulation study shows the relation between the amount of prior information and results in terms of estimation and the non-null detection rate. With this information, researchers can observe the relation between the specificity of prior information and other factors such as estimation problems, bias, non-null detection rate, and cover- age. This paper is not a demonstration of how prior distributions should be manipulated to secure statistically significant results: This would not be an ethical use of the information, and the exact results may vary between study variables and models. As shown in Zondervan-Zwijnenburg et al. (2017a) , prior knowledge has to be acquired systematically and specifications of prior distributions have to be justified. Moreover, to promote transparency, we advise to demonstrate the impact of other priors on the results by means of a sensitivity analysis (see also Depaoli and Van de Schoot, 2017). We believe that the manipulation of priors to obtain a "desirable" result would fall under unethical research practices.

Another cautionary note should be made on the use of default priors for variance parameters with small samples. Variance and disturbance parameters were not the focus of this study, but it has been shown, for example, by McNeish (2016a) and Van de Schoot et al. (2015) that these estimates can be severely biased with uninformative priors.

2.6.2 Final Recommendations

Based on these findings, we recommend researchers with focal groups with fewer than 200 participants to conduct a simulation study in order to evaluate the impact of the small sample on estimation issues, bias, coverage, and non-null detection rate. When maximum likelihood estimation cannot generate proper output under the circumstances of interest, we suggest to obtain prior information. Zondervan-Zwijnenburg et al. (2017a) provides guidelines on collecting and including prior information. If sufficiently precise prior information can be acquired, the data can be analyzed. If the researcher is not able to meet the requirements, simpler models (e.g., descriptive statistics, case studies), waiting until more prior information and participants become available (e.g., by following Google Scholar Alerts, RSS feeds, and reapproaching schools in a new academic year), or conducting the analysis to contribute to cumulative science without making inferences, are alternative ways to deal with the data.

3

3rd

Where do priors come from? Applying guidelines to construct informative priors in small sample research

Summary. This article demonstrates the usefulness of Bayesian estimation with small samples. In Bayesian estimation, prior information can be included, which increases the precision of the posterior distribution. The posterior distribution reflects likely parameter values given the current state of knowledge. An issue that has received little attention, however, is the acquisition of prior information. This study provides general guidelines to collect prior knowledge and formalize it in prior distributions. Moreover, this study demonstrates with an empirical application how prior knowledge can be acquired systematically. The article closes with a discussion that also warns against the misuse of prior information.

Small samples occur regularly in social sciences for various reasons. Sometimes the size of the population is extremely limited, for example in children with a rare disease (Van der Lee et al., 2008), or juvenile females charged with murder (Roe-Sepowitz, 2009). The population can also be difficult to recruit and prone to drop-out, because they are homeless, institutionalized, or playing truant (Mäkelä and Huhtanen, 2010; McCabe et al., 2016; Peeters et al., 2014). Factors such as costs (Rocchetti et al., 2013) and ethical constraints (Van der Lee et al., 2008) may also make efforts to obtain a larger sample quite difficult (or impossible).

One of the consequences of small samples such as those described above is low statistical power (i.e., inflated Type II error, see for example Muthén and Curran 1997 for a simulation study). Non-significant p-values, which likely follow from underpowered analyses, cannot be meaningfully interpreted in the null hypothesis significance testing

This chapter is published as Zondervan-Zwijnenburg, M.A.J., Peeters, M., Depaoli, S., & Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, *14*(4), 305-320. doi: 10.1080/15427609.2017.1370966

Author contributions: MZ and RS designed the study. MP collected the data. MZ wrote the paper under supervision of RS. Additional feedback was provided by SD and MP.

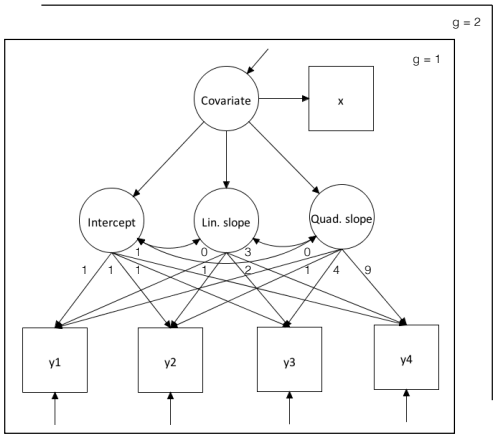


Fig. 3.1: Multiple group latent growth model with one covariate and groups indicated by g. y_1^g, y_2^g, y_3^g, y_4^g represent four assessments of a developing construct with residual error variances. x^g is a time-invariant predictor of growth that represents the latent variable Covariate^g without measurement error. The regressions of the latent growth factors Intercept^g, Lin. slope^g, and Quad. slope on the Covariate^g are equal over groups.

(NHST) framework. Consequently, researchers often refrain from analyzing interesting (i.e, exceptional) groups, and deviate from recommended cut-offs to cover larger groups (e.g., Heron et al., 2013; Scharkow et al., 2014), or need to conclude that power was too low to detect the foreseeable small effect (e.g., Mahu et al., 2015).

In Bayesian statistics, on the other hand, prior information is combined with the data in the analysis resulting in a posterior distribution that, irrespective of sample size, can be interpreted as a distribution displaying the probability of parameter values. Prior distributions can incorporate information about model parameters that researchers have before seeing the data. Sometimes, researchers lack information, but often they are able to limit the admissible parameter space. For example, a prior distribution for a mean could exclude values that are outside the range of the measurement scale. Such a specification already increases the precision of the posterior distribution. Alternatively, a researcher may be able to specify a normal prior distribution that favors some values over others. When no prior information is available, an uninformative prior can be adopted which typically specifies a wide range of parameter values as probable. When a prior distribution becomes narrower, because more prior information becomes available, the posterior distribution is affected increasingly by the prior information and becomes narrower and more informative as well. One could state that statistical power increases, and inference errors are less likely to occur (see Van de Schoot et al.

To readers interested in a gentle introduction into Bayesian statistics for social scientists, we recommend Kruschke (2014), and Van de Schoot et al. (2013).

The term *power* originates from the frequentist setting, where only frequency probabilities can be considered. A frequency probability refers to the expected relative frequency of an

2017 for an overview of simulation results in the last 25 years). Furthermore, prior distributions can avoid inadmissible estimates and convergence issues in Bayesian estimation. Typical frequentist estimation methods like maximum likelihood (ML) estimation have been shown to suffer from these problems by several simulation studies (e.g., Boomsma and Hoogland, 2001; Hox and Maas, 2001; Meuleman and Billiet, 2009; Tolvanen, 2000). In sum, Bayesian estimation with informative priors results in meaningful output, even with small samples, and can increase the precision of the result when prior information is available. Note however, that even though prior information will increase the precision of the estimates with small samples, we strongly recommend collecting larger samples if this is possible in any way.

The advantages of Bayesian estimation with informative priors for small sample research may be clear, but the current literature does not demonstrate how prior information can be collected systematically, and how priors subsequently should be specified with the obtained information. The current study addresses this gap in the literature. First, we present guidelines to support researchers that are interested in conducting a Bayesian analysis with informative priors. However, we do not only present guidelines, we also report on our efforts to actually follow them in the context of an empirical application concerning a latent growth model. For this application, we search for prior information, and subsequently formalize the information into prior distributions. The empirical application does not represent the "ideal" situation. On the contrary, the application is a realistic study for which prior information is not easily acquired. With this application, we show what a researcher can do to obtain prior information in complex situations, and what there is left to do, when things do not work out as hoped for. We expect that social scientists who happen to operate under ideal circumstances can easily derive the appropriate steps to take from this example as well. Finally, the paper provides a discussion on how priors should, and should not be used.

3.1 Guidelines

The following general guidelines can support researchers interested in constructing informative priors for parameters:

- Determine what strategy suits the project of interest best with questions like:
  - Could prior information likely be found in the literature (e.g., meta-analyses, reviews, empirical studies)? Note that the quantification of prior information is

outcome given repeated events. Power is the frequency probability of rejecting the null, given that the alternative hypothesis is true. In the Bayesian setting, usually subjective probabilities reflecting a degree of belief are considered, but a subjective probability can be translated to a frequency probability (Press, 2009). For example, a probability of .5 of getting heads from a coin flip can be translated to 5 expected heads in 10 tosses. Hence, one can imagine that a construct like power can be used in the Bayesian context as well (see also Rubin, 1984).

more straightforward when the literature covers the same variables obtained with the same measures as the data of interest.
  - Are there experts on the subject matter, and who are they? How can experts contribute? Would experts be able to specify priors for the parameters in the model at hand, or can they contribute in a different manner?
  - What general knowledge is available about the model parameters?
  - Is it possible to increase the information in the data by increasing the sample size?
- Determine how to gather the information systematically. Keep a log of every decision (see, for example, the logbook provided at `osf.io/aw8fy`).
- When you intend to construct informative priors, visualize them. A visualization (e.g., with R, or `www.wolframalpha.com`) quickly shows whether the prior specifications that you consider are reasonable.
- When conducting a Bayesian analysis, always provide the following: (1) the origin of and reason behind the priors, and (2) the exact specifications of the priors. See Depaoli and Van de Schoot (2017) for further instructions on reporting Bayesian analyses.
- Conduct a sensitivity analysis and show the impact of various priors on the posterior estimates (Van de Schoot et al., 2017). Consider at least the derived informative priors and default priors, but conservative or skeptical priors may be interesting to examine as well.
- Try to understand and interpret differences between analyses with different priors.

3.2 Empirical Application

To demonstrate how prior information can be systematically collected and included in a Bayesian analysis, we compared the development rate of cognitive performance in young heavy cannabis users to that of their nonusing peers in a two-group latent growth model (LGM). We did so in a high-risk sample of young adolescents enrolled in special education because of behavioral problems (Peeters et al., 2014). *Young* was defined as younger than 15, because cannabis use before age 15 is considered as early onset (Jacobus et al., 2009). A relation between cannabis use and poorer attention, learning, and processing speed is expected especially with early onset of use (Fontes et al., 2011; Jacobus et al., 2009; Schweinsburg et al., 2008). By using a high-risk sample, the heavy cannabis users and their nonusing peers are better comparable. However, this also limits the total sample size. In addition, heavy cannabis using adolescents were expected to be a minority even in this sample. Thus we have a small and unbalanced samples, for which several simulation studies have demonstrated that ML estimation results in low power, and computational issues (e.g., Hox and Maas, 2001; Meuleman and Billiet, 2009; Muthén and Curran, 1997).

In Bayesian estimation informative priors can increase the precision of the posterior outcome, and even when the statistical power would be low, the posterior distribution would still be meaningful and easy to interpret. From the posterior distribution, a

measure of central tendency (e.g., the mean, median, or mode) is usually taken to reflect a point estimate for the parameter of interest. Additionally, a 95% credibility interval can be derived from the posterior distribution. This interval has a 95% chance of containing the true parameter value, given the data and the prior. The frequentist 95% confidence interval, in contrast, cannot be interpreted as an interval that has a 95% chance of containing the true value. The confidence interval only contains the true population value in 95% of the intervals over a long run of trials. Thus, the Bayesian framework provides solutions that are more meaningful. Additionally, prior distributions can prevent inadmissible solutions by assigning zero probability to ranges of values that the parameter cannot take (e.g., negative values for variances). All in all, we had various reasons to conduct a Bayesian analysis and search for prior information.

3.2.1 Method

Participants

The original study of Peeters et al. (2014) concerned 374 adolescents (330 boys, 44 girls) who attended special education schools for youth with externalizing behavioral problems in the Netherlands. From this group, adolescents younger than 15 at the first assessment were selected to ensure that cannabis use at the first wave reflected an early onset. Twenty-eight participants did not indicate their age in years at the first wave. To avoid a loss of power, missing data for age was imputed by means of the R-package `mice` (Van Buuren and Groothuis-Oudshoorn, 2011). Participants' ages could be easily imputed, because age in full years was assessed repeatedly in the two years that assessments were taken. Exact birth dates, however, were not available. The mean age over 10 imputations for each participant was computed. Participants with a rounded mean age younger than 15 were selected for further analyses ($n = 331$).

Subsequently, we mimicked previous literature (Mahmood et al., 2010) in that non-users and heavy users were selected to contrast the two extremes. Students were selected based on their response to the question: "How often have you used cannabis during the past 6 months?". The five answer categories to this question were: (1) "I have not used cannabis/marijuana", (2) "Once a month", (3) "2-4 times a month", (4) "2-3 times a week", and (5) "4 times a week or more". Adolescents who selected the first answer category "I have not used cannabis/marijuana" were identified as non-users ($n = 252$, mean age = 13.30, 90.4% male). All adolescents who selected the fourth and fifth answer category ($n = 16$, mean age = 13.38, 81.3% male) met the requirements to be considered heavy cannabis users (Barnes et al., 2005). The 25 and 13 participants that selected category 2 and 3 respectively were not included, as well as the 25 participants that chose not to answer this question.

Measures

Working memory. Working memory performance was selected as a measure of cognitive performance because working memory continues to develop throughout adolescence

(Best and Miller, 2010). Working memory performance was assessed with the non-verbal self-ordered pointing task (SOPT) with representational drawings of everyday objects (Petrides and Milner, 1982). In this task, participants were instructed to select a different picture out of a set of pictures each time, while after each choice the location of the pictures changed and they were not allowed to select the same location consecutively. The task included one practice trial with a set of 4 unique pictures, and four assessment trials with sets of 6, 8, 10, and 12 unique pictures. The percentage of correct choices on the task was used as an indication of working memory performance. Details of the assessment can be found in Peeters et al. (2014). In the current dataset, working memory performance was assessed four times over two years with intervals of approximately 6 months (Peeters et al., 2014).

Alcohol use. We corrected the development of both groups for the impact of quantity and frequency of alcohol use at the start of the study, as recommended by Jacobus et al. (2009). Alcohol use was assessed by means of a quantity frequency measure (QF). The QF was a multiplication of the number of days a week that the adolescent usually consumed alcohol with the number of glasses that were usually consumed on drinking days. A detailed description can be found in Peeters et al. (2014).

Statistical Approach

To investigate the difference in cognitive development between heavy cannabis users and non-users, the latent growth model as shown in Figure 3.1 was the preferred analysis. The repeated measures (i.e., y_1^g, y_2^g, y_3^g, and y_4^g) were represented by the four assessments of SOPT scores, and the covariate for this model was a measure of alcohol use quantity and frequency at the start of the study (i.e., x^g). The quadratic slope was included, because the linear increase in the percentage of correct responses on the task was expected to level off over time. Because this effect was expected to be similar for both groups, the quadratic slope was constrained equal over groups accordingly. The linear growth factor in this model represents the linear growth rate at the first time point while a quadratic factor is modeled. As indicated above, the model has one covariate representing an observed time-invariant predictor. As a result the latent time variables technically have intercepts instead of means. However, to avoid confusion between the intercept growth factor and the intercepts of the latent growth factors, the latter will be referred to as means throughout the article. To assess the growth rate difference between groups, a new parameter (denoted by $\Delta\alpha$) was constructed by subtracting the linear slope mean of the frequent users group (i.e., the exceptional group) from that of the non-users group (i.e., the reference group).

3.2.2 Prior knowledge

Prior distributions need to be specified for all parameters in a Bayesian model, but we focused on finding prior knowledge for our main parameters: $\Delta\alpha$, and the latent growth means. For the remaining parameters, we used the default settings of Mplus 7.3, that is: $N(0, 10^{10})$ for mean of the covariate and for the regression coefficients,

$IG(-1, 0)$ for the residual variances and the variance of the covariate, and $IW(0, -4)$ for the variance-covariance matrix of the growth factors (Muthén and Muthén, 2012). Note that default settings can cause problematic results. See for instance Van de Schoot et al. (2015). We report a sensitivity analysis with varying prior distributions for the remaining parameters in our logbook, which is provided at `osf.io/aw8fy`.

Prior knowledge can be extracted from several resources such as meta analyses, reviews, empirical studies, and experts (O'Hagan et al., 2006). We evaluated potential sources of prior information one by one, and after consideration of each source, it was re-evaluated what the next step would be.

Meta-Analyses

A literature search was conducted in Scopus for meta analyses published between January 2000 and December 2013 based on the terms: *cannabis*, *marijuana*, *adolescent*, and *cognitive*. The search yielded six results. However, none of them were relevant because they concerned non-healthy subjects (i.e., suffering from psychosis, or schizophrenia; $n = 3$), and interventions ($n = 3$), instead of the relation between cannabis use and cognitive impairment (see `osf.io/aw8fy` for references). As a result, the search for prior information had to be continued with respect to the next source: Reviews.

Reviews

A search for reviews with the same keywords as for meta analyses yielded 33 English matches. We had to exclude 27 of these studies, because they concerned preventions and interventions ($n = 11$), schizophrenia and substance use disorders ($n = 4$), prenatal exposure ($n = 5$), or did not focus on cognitive effects of cannabis use ($n = 7$). Consequently, six reviews were considered relevant. Three additional relevant reviews were identified through other resources. The resulting nine reviews were all published in 2008 and 2009, and covered information from 36 articles, including human and animal (preclinical) studies. By analyzing key sentences from the reviews, we learned that a zero effect size for $\Delta\alpha$ should receive more than zero probability from the priors (see `osf.io/aw8fy` for references and details). Quantitative information about the exact values of the intercept, linear, and quadratic slopes, however, lacked. To find this information with which priors can be constructed, we decided to continue with a search for actual SOPT scores in empirical articles.

Empirical Studies

Because it is not common to mention an assessment instrument in the title, abstract, or keywords of an article, a search engine that evaluates the content of complete articles had to be used. A suitable search engine for this purpose is Google Scholar. In Google Scholar, we used the following search query: "self-ordered pointing", child OR adolescent. The search yielded 693 hits. To obtain the most relevant results for

our research population, several inclusion criteria were applied. First, actual scores of the SOPT with familiar objects had to be provided in the study. Second, the mean age of the samples studied had to be between 9.5 and 17.5 years old, this age range covers the age of the research population ± 4 years. Third, the version of the SOPT had to include concrete pictures, because other versions differ in difficulty, and thus in their scores. Fourth, samples had to consider typically developing children, or children with attention deficit/hyperactivity disorder (ADHD), oppositional defiant disorder (ODD), and/or conduct disorder (CD). ADHD, ODD, and CD are disorders commonly encountered in special education classes such as those included in the current study. Fifth, studies had to cover samples that were not already covered in (1) previous articles that met the inclusion criteria or (2) the current study.

After correspondence with authors about task and sample ambiguities, 13 out of 693 articles yielded useful information. All obtained SOPT scores were transformed into a percentage of correct responses. An overview of the articles with encountered SOPT scores for children and adolescents is given at `osf.io/aw8fy`. To ensure that the obtained scores were relevant for our specific high-risk sample, we involved experts.

Experts

Two experts were recruited to participate in the current study: A developmental psychopathology professor and a clinician at a secondary school for youth with externalizing behavioral problems. In separate face-to-face meetings, the experts received a questionnaire consisting of an explanatory text and a table. Based on sample descriptions from the selected empirical studies, the experts rated the relevance of these samples for the population of youth with behavioral problems in general, and they estimated the percentage of cannabis users in the described sample. During the procedure, the experts did not get information on the SOPT scores in the study, nor did they get information about the authors of the study. The intraclass correlation coefficient with respect to the absolute agreement of the two experts about study representativeness was .87, indicating good interrater reliability.

The relevance of the samples rated by both experts was averaged. When the average judgment of sample relevance was higher than .5, the sample relevance was multiplied with the sample size, resulting in a number that was interpreted as the relevant sample size. Based on the relevant sample sizes, a weighted average of the SOPT scores for each age group was computed. Relevant samples with an estimated percentage of cannabis users higher than 50% were considered relevant for the exceptional group.

Figure 3.2 shows the weighted averages by age and population. As can be seen, only one sample qualified as representative for the exceptional population of heavy cannabis-using youth with externalizing behavioral problems according to the experts. Four samples were considered relevant for the reference group. However, these studies all covered 10-year-olds, yielding only one datapoint from a longitudinal perspective. To construct a prior for the intercept factor at age 13 and the linear slope factor, prior information had to be obtained for at least two age groups. Because these were not

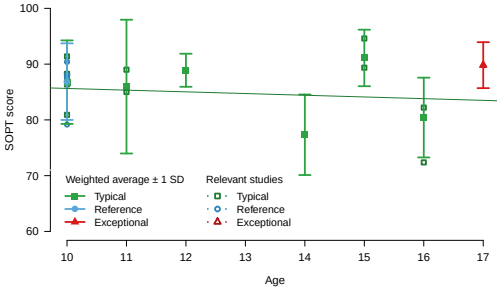


Fig. 3.2: Weighted SOPT scores by age.

available for the population of interest, general knowledge needed to complement the information that we had acquired so far.

General knowledge and prior specification

Intercept. The information on typically developing children indicated that around age 13.5 values between 75 and 95 are most likely. Hence, the prior mean was set at the mean of those values: 85. To determine the variance for this mean, prior mean variances were visualized (see Figure 3.3a) and the preference of the distribution for some values over others was calculated. With respect to the intercept factor, a variance of 30 implied that values between 80 and 90 were 1.06 times as likely in the prior distribution as values between 90 and 100 (or between 70 and 80), and 1.25 times as likely as values between 100 and 110 (or between 60 and 70). These ratios in the likelihood of values in the prior distribution were considered reasonable, and thus implemented as such.

Linear Slope. To acquire an idea about the trend over time, a linear regression was fitted to the SOPT scores for typically developing children. The result is represented by the slope in Figure 3.2. The negative trend, however, is not in line with validated theory (Best and Miller, 2010). In addition, the SOPT scores of typically developing children seemed inconsistent over time. Based on theory, we expected a positive development of SOPT scores over time (Best and Miller, 2010). This expectation was confirmed by the empirical study of Clarke (2009), who found a significant positive cross-sectional correlation of medium size between age and performance on the SOPT for children with ADHD, who were at risk for CD. However, because the prior information derived from the SOPT scores for typically developing children indicated a negative trend, negative values were not excluded. More specifically, given an intercept prior mean of 85, the linear growth mean could be up to 1.75 points per 6 months to reach a score of 95.5 at age 17. However, we expected that the growth rate decreased when higher scores were achieved. Thus, a negative quadratic factor was anticipated, which allowed for a higher linear growth factor. Given the expectation of a negative quadratic factor

mean, the linear slope prior mean was set at 2.00 points per 6 months with a prior variance of 7.5. This prior variances caused values between 0 and 4 to be 1.15 times as likely in the prior distribution as values between 4 and 8 (or between 0 and -4). See Figure 3.3b for a visualization.

Quadratic Slope. As mentioned, a diminishing growth rate over time can be represented by a negative quadratic growth factor. In the current empirical application, however, a large negative quadratic growth factor might cause a decrease in working memory within the range of the model, whereas this is not expected during adolescence (Best and Miller, 2010). Therefore, the prior mean for the quadratic growth factor was set at -0.1 with a variance of 7.5. The variance of 7.5 is relatively wide for this quadratic slope that we expect to be small. With this variance we reflect that we do not have very specific information for the quadratic slope factor. This distribution has the same shape as that in 3.3b, but is shifted 2.1 points to the left. The combination of the specified priors for the latent growth factor means presupposed an increase of 5.1 in the percentage of correct SOPT entries over the two years in which data was collected.

All in all, the final prior distributions were:

$$p(\text{Intercept}^g) \sim N(85.0, 30.0), \tag{3.1}$$

$$p(\text{Lin. slope}^g) \sim N(2.0, 7.5), \tag{3.2}$$

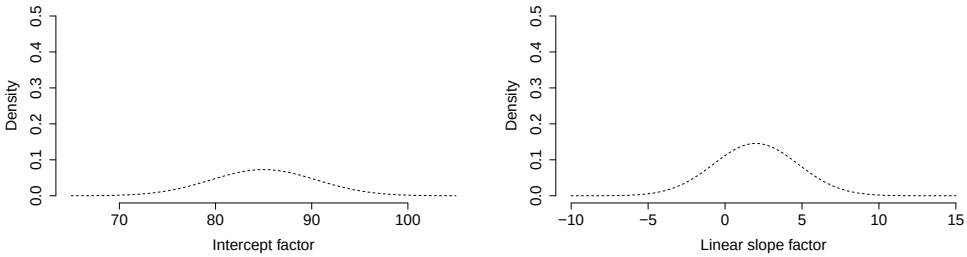
$$p(\text{Quad. slope}) \sim N(-0.1, 7.5). \tag{3.3}$$

Note that we specified normal distributions in all cases, but that other distributional forms (e.g., beta, Cauchy, skewed normal, etc.) can also be considered. Novice appliers of Bayesian statistics might need to be aware of software limitations in this respect. See Depaoli and Van de Schoot (2017) for detailed guidelines on specifying prior distributions.

3.2.3 Results

A Bayesian analysis was conducted in Mplus 7.3 with four chains, a minimum of 50,000 iterations, and BCONVERGENCE was set at an extra strict number of .005. BCONVERGENCE affects the pursued Gelman-Rubin potential scale reduction (PSR) (Gelman and Rubin, 1992) criterion value for the model to be considered converged (Muthén and Muthén, 2012). Convergence was obtained at 50.000 iterations. Subsequently, the first half of the iterations was discarded as a burn-in phase. The maximum PSR among the iterations that contributed to the posterior results (i.e., 25.000-50.000) was 1.014. The median of the posterior distribution was interpreted as

No prior was assigned directly to $\Delta\alpha$, since this parameter is derived from the linear slope means. To implement a difference between groups with a small effect size, as was indicated by the reviews, information about the residual variance in the linear slope after prediction by the amount of alcohol use was necessary. This information could not be derived from any of the evaluated literature.



(a) $N(85.0, 30.0)$ (b) $N(2.0, 7.5)$

Fig. 3.3: Visualizations of the prior distributions for the latent growth factor means.

the point estimate. Traceplots for all parameters showed that the chains had stable means and variances. See `osf.io/aw8fy` for the data, syntax, and output.

The results of the analysis are provided in Table 3.1. The 95% (highest posterior density) interval for $\Delta\alpha = [0.17, 7.30]$ (see also Figure 3.4). The median of this distribution is 3.77. Based on this distribution we can state that we are 98.0% sure that $\Delta\alpha > 0$. Thus, young adolescents not using cannabis seem to have a higher working memory increase than their heavy cannabis using peers. Cohen's d at the median of this distribution is 0.54.

To evaluate the impact of the informative priors on this result, a sensitivity analysis was conducted, which is presented in detail at `osf.io/aw8fy`. The main results for the analysis with default priors are presented in Table 3.1. The 95% credibility intervals for all parameters in the analysis with informative priors were smaller than the analysis with default priors, indicating that the prior information increased the precision of the final results. The last column shows the relative difference in posterior medians between both analyses. In absolute terms, the discrepancies ranged from 1.07% for the non-users' intercept to 51.53% for the heavy users' linear slope mean. The relative difference between the analyses for $\Delta\alpha$ was 46.48%. In the analysis with default priors, the 95% interval was $\Delta\alpha = [-2.05, 7.28]$ with 2.57 as its median (Cohen's $d = .35$). 86.4% of the posterior distribution for $\Delta\alpha$ is larger than 0, and 68.4% is in line with at least a small effect size (i.e., Cohen's $d \geq .20$). The posterior distributions for $\Delta\alpha$ obtained from the analysis with informative priors and the analysis with default priors are depicted in Figure 3.4. The figure shows that relative to the posterior from the analysis with informative priors, the posterior from the analysis with default priors is wider and puts a higher probability on lower values for $\Delta\alpha$.

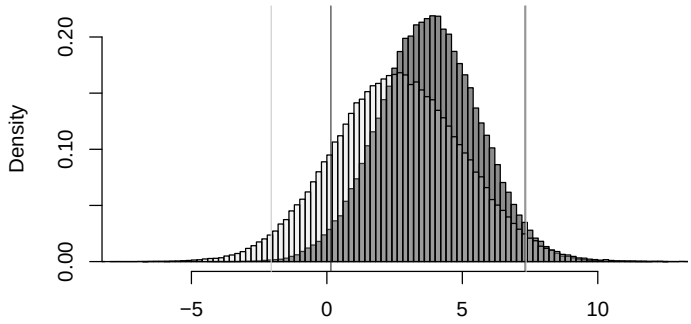


Fig. 3.4: Samples from the posterior distributions for $\Delta\alpha$ based on informative priors (darkgrey) and default priors (transparant). Vertical lines indicate the limits of the associated 95% highest posterior density interval.

Table 3.1: Main posterior parameter estimates for the analysis with informative and default priors.

	Informative priors	Default priors	Difference
Parameter	M 95% CI	M 95% CI	%
Intercept$_{non\text{-}users}$	74.00 [72.47, 75.55]	73.22 [71.58, 74.83]	1.07
Intercept$_{users}$	78.68 [73.06, 84.20]	75.16 [68.18, 82.30]	4.68
Lin. slope$_{non\text{-}users}$	6.10 [3.94, 8.26]	7.40 [4.95, 9.95]	-17.68
Lin. slope$_{users}$	2.34 [-1.28, 5.96]	4.83 [-0.35, 10.00]	-51.53
Quad. slope	-1.58 [-2.27,-0.92]	-1.94 [-2.72,-1.18]	18.48
$\Delta\alpha$	3.77 [0.17, 7.30]	2.57 [-2.05, 7.28]	46.48

Note. $n_{non\text{-}users} = 252$, $n_{users} = 16$. CI = credibility interval.

3.2.4 Conclusion Empirical Application

Despite the expected lack of statistical power, the analysis did show that young adolescents not using cannabis most presumably have a stronger working memory growth rate than their heavy cannabis-using peers. Note that we cannot draw causal conclusions, because there may be more differences between the sample of heavy cannabis users and non-users that relate to their working memory development rate than cannabis alone, even after controlling for quantity and frequency of alcohol use at the start of the study. Furthermore, the information from the data for both groups was substantiated with more general prior information. This general information

affected the posterior distribution for the heavy cannabis users more, because the sample for this group was smaller and the data thus contained less information. The sensitivity analysis showed that with default priors we would also conclude that we expect that young adolescents not using cannabis most presumably have a stronger working memory growth rate than their heavy cannabis using peers, but we are less confident. Future studies can use our posterior results, either with informative or default priors, as prior information again. By reporting the exact prior distributions and how they came into being, everyone can review our prior information. By reporting an analysis with uninformative priors as well, we provide a similar insight into data.

3.3 Discussion

The aim of the current study was to provide guidance on and to demonstrate how prior information can be collected systematically and subsequently formalized, since this information is lacking in current literature. In the pursuit of this aim, we provided guidelines and demonstrated their application with an empirical application. In the following paragraphs we discuss advantages and disadvantages of different prior information sources, we discuss directions for future research, we discuss the ethical use of Bayes, and end with some concluding thoughts relating back to NHST.

When prior information is scarce, it seems promising to collaborate closely with experts. In the current study, experts contributed to the evaluation of information obtained from other studies. Another option is to let experts determine the prior for the parameter of interest themselves. In that case, the researcher must ensure that the experts understand the parameter of interest, use appropriate heuristics, and avoid fallacies (see also O'Hagan et al., 2006). Under these conditions, collaborating with experts can always increase the precision of the result, in contrast to searching for literature, which may not result in useful prior information. Additionally, published studies may suffer from publication bias. It is important to realize however, that "academic" experts may be affected by this publication bias as well. Furthermore, a procedure to elicit priors for the specific model parameter adjusted to the experts at hand may be nonexistent. Developing a valid and reliable procedure to elicit prior information may be a full research project in itself (see for example Johnson et al., 2010b; Zondervan-Zwijnenburg et al., 2017b), whereas a search for prior information in the literature may resemble an extended systematic literature study that researchers would also conduct to write the introduction to their paper. Currently, experts in the social sciences mainly contribute to clinical studies by estimating (success) probabilities (Spiegelhalter et al., 2000). Empirical research on the elicitation of more complex parameters within the social sciences is warranted (O'Hagan et al., 2006). As was also apparent in the empirical application, prior information does not only affect standard errors, it can also change estimates in case of a discrepancy between the prior information and the data. In the analysis with informative priors, the posterior results in the exceptional group were affected more by the prior distributions than the posterior results of the reference group. The reference group posterior distributions

were mainly affected by the data. Additional research is required with respect to the inclusion of prior information. More specifically, the area of research about the inclusion of highly informative prior distributions is still in its early stages. Researchers may want to test whether a mismatch between the prior and data exists. Methods for such a test need to be further developed to be applicable for applied researchers with all sorts of models. Additionally, some of the reviews discussed in the empirical application considered information from animal studies, but how well can this information serve as prior information in social and behavioral sciences research, should it be merged with prior information from studies on humans, and if so how?

3.3.1 (Un)ethical use of Bayesian estimation

Like frequentist NHST, Bayesian estimation methods can be misused. Misuse of Bayesian estimation with informative priors would be to repeatedly conduct analyses with varying priors and only report the analysis with "desirable" results. This is unethical behavior, comparable to '*p*-hacking' and data fabrication.

Instead, researchers should be transparant about the actions and reasoning that led to the priors at hand. In the current study, for example, we conducted a systematic search, we reported this search, and provided justifications for the final prior choices. In this manner, readers can decide for themselves whether they are convinced by the information.

A simulation study can clarify how specific prior information should be to obtain posterior results that can convincingly exclude specific parameter values like zero. This may be helpful in designing the search for prior information. However, if the results show that zero will be a likely value a posteriori, researchers should be able to accept this as a conclusion. In studies that are conducted properly, such results should be regarded publication worthy. Irrespective of the results, any publication can provide prior information for future studies on the same topic. In this manner, cumulative science through Bayesian updating is promoted.

Additionally, to promote transparency we advise to demonstrate the impact of other priors on the results by means of a sensitivity analysis (Van Erp et al., 2018). The sensitivity analysis should be clearly documented as well (see, for example, the logbook provided at `osf.io/aw8fy`). Clear reporting and sensitivity analyses contribute to transparency, and thus integrity, that is recognized to be important for the survival of social science research (Cumming, 2014). Depaoli and Van de Schoot (2017) developed a 10-point checklist to improve transparency and replication in Bayesian research.

3.3.2 Concluding Thoughts

The issues with NHST have been widely discussed (e.g., Cumming, 2014; Cohen, 1994; Kline et al., 2004; Rozeboom, 1960), and the Bayesian framework offers a viable alternative to this hypothesis testing framework, because it can prevent researchers from having to make an over-simplified decision of whether a hypothesis is to be rejected. Bayesian estimation is a beneficial tool that is less restrictive than the

conventional NHST framework. It is our hope that this demonstration of how informed priors can be acquired and implemented will aid in broadening the methods typically used for assessing hypotheses in the conventional framework.

The current study showed how prior information can be obtained systematically, and how this information can be formalized into prior distributions. Once again we want to emphasize that specifying highly informative prior distributions is not to be used in order to achieve statistically significant results. Instead, prior specifications should be used because including available information can be the key to answering questions about populations that otherwise remain unanswered. The search for prior information may be intensive and time consuming, yet it can be rewarding because it provides great insight in the current state of the field, it can improve the analysis, and it results in an update of knowledge.

4

Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations

4th

Summary. The purpose of the current study was to apply and evaluate a procedure to elicit expert judgments about correlations, and to update this information with empirical data. The result is a face-to-face group elicitation procedure with as its central element a trial roulette question that elicits experts' judgments expressed as distributions. During the elicitation procedure, a concordance probability question was used to provide feedback to the experts on their judgments.

We evaluated the elicitation procedure in terms of validity and reliability by means of an application with a small sample of experts. Validity means that the elicited distributions accurately represent the experts' judgments. Reliability concerns the consistency of the elicited judgments over time. Four behavioral scientists provided their judgments with respect to the correlation between cognitive potential and academic performance for two separate populations enrolled at a specific school in the Netherlands that provides special education to youth with severe behavioral problems: youth with autism spectrum disorder (ASD), and youth with diagnoses other than ASD. Measures of face-validity, feasibility, convergent validity, coherence, and intra-rater reliability showed promising results.

Furthermore, the current study illustrates the use of the elicitation procedure and elicited distributions in a social science application. The elicited distributions were used as a prior for the correlation, and updated with data for both populations collected at the school of interest.

This chapter is published as Zondervan-Zwijnenburg, M.A.J., Van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & Van de Schoot, R. (2017). Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations. *Frontiers in Psychology*, *8*, 90. doi: 10.3389/fpsyg.2017.00090

Author contributions: MZ and RS conceptualized the study. Development of the expert elicitation questionnaire was performed by MZ, RS, WH, and KL. The pilot test of the expert elicitation questionnaire was directed by MZ with support from KL, and facilitated by RS. Expert elicitation was performed by MZ, and facilitated by WH. Data was analyzed and stored by MZ. HH proposed, among other things, to pool posterior distributions as an alternative of updating the pooled prior, and closely monitored Appendix A.2 and A.6. MZ wrote and revised the paper with feedback of RS and HH.

The current study shows that the newly developed elicitation procedure combining the trial roulette method with the elicitation of correlations is a promising tool, and that the results of the procedure are useful as prior information in a Bayesian analysis.

> "Expert judgement has always played a large role in science and engineering. Increasingly, expert judgment is recognized as just another type of scientific data ..." (Goossens et al., 2008, p. 236).

This quote is the result of 15 years of developing and applying expert judgment elicitation procedures at TU Delft in the Netherlands. In the sectors of, for example, nuclear applications, chemical industries, water pollution, volcano eruptions, space shuttles, aviation, health, banking, and occupational hazards over 800 experts have conducted elicitations on over 4000 variables (Goossens et al., 2008). In social science, however, expert judgment is seldom used for estimation and inference, especially not in combination with data (see Spiegelhalter et al. 2000 and O'Hagan et al. 2006 for a few examples in health care). This may be partly explained by the fact that the Bayesian framework that allows for the inclusion of prior knowledge elicited from experts in data analyses was adopted much earlier and on a far greater scale by fields of science, technology, engineering, and mathematics as compared to social science, arts, and humanities (Van de Schoot et al., 2017). Nevertheless, the use of Bayesian statistics is increasing in social science as well.

In Bayesian statistics, a prior distribution containing probable values for each parameter of a model is updated with data, resulting in a posterior distribution: an updated summary of the knowledge about the model parameters. Expert judgments can be a useful source of prior information, especially when data is scarce (Hampson et al., 2014). Small samples contain a limited amount of information, and the reliability of the data may be questionable. Expert judgments can complement the information from the data. Additionally, updating current expert judgments with new data can also be a research goal in itself. The updated result can increase confidence in original views of experts, or adapt these views. In the current study, we focus on the elicitation of a correlation between two variables. The correlation–our key parameter –is modeled in a bivariate normal distribution that consist of two means, and two standard deviations next to the correlation parameter itself. Figure 4.1 shows the research cycle that can be followed when expert judgments for a key parameter are to be updated with data.

When the research objective is to update expert judgments with current data, these judgments have to be elicited first (see Figure 4.1, step 2). The elicitation of judgments is a sensitive process, because the human mind tends to employ easy-to-use strategies that are not necessarily rational or optimal (O'Hagan et al., 2006; Van Lenthe, 1993). The elicitation of correlations between variables has received considerable attention in fields other than social science. Kraan (2002) and O'Hagan et al. (2006) describe, for example, (1) a method where strength of the relationship between variables is expressed on a 7-point Likert scale, (2) a method where the expert is requested to

For an introduction to Bayesian statistics for social scientists we recommend Gill (2014), Kaplan (2014), and Van de Schoot et al. (2013)

Fig. 4.1: Research cycle to update expert judgments with new data.

provide Spearmans's correlation, (3) a method where the probability of concordance is assessed (further explained in a later section), and (4) a method that requests conditional quantile estimates. Clemen et al. (2000) evaluated six methods to elicit judgments about correlations with respect to accuracy, variation among experts, and difficulty. The best method according to their study was to simply ask experts to report a correlation. However, many others are critical to the capability of the human mind to assess a correlation (Gokhale and Press, 1982; O'Hagan et al., 2006; Morgan et al., 1992). It is clear that determining a correlation is not an easy task. Hence, instead of eliciting a point estimate as in the above methods, we consider it important to elicit a full distribution that captures the experts' uncertainty as well.

One way to elicit continuous distributions is to ask the expert to specify fractiles or quantiles of the distribution of interest such as the 5th, 50th and 95th. After a training with respect to quantiles, a question to obtain the 5th percentile for the mean of IQ in a specific population may be: "Can you determine a value such that the mean of IQ is 5% likely to be less than this point and 95% likely to be greater than this point?" (O'Hagan et al., 2006). Such a question should be asked for all desired quantiles. Alternatively, one could ask for multiple quantiles at once, for example: "To capture your uncertainty please provide the 5th, 25th, 50eth, 75th and 95th percentiles of your uncertainty distribution" (Morales Nápoles, 2010, p. 82). Morales Nápoles (2010) used this method to elicit a distribution for a correlation. After the elicitation phase, distributions are fitted to the elicited quantiles (Cooke and Goossens, 1999).

Another way to obtain uncertainty distributions is the trial roulette method (Gore, 1987). Experts are provided with a number of "chips" to allocate probability to bins of a histogram (see Figure 4.2). With 20 chips, each chip represents five percent probability. The number of chips placed over a certain value reflects how probable the value is according to the expert. Several variants on this method have been developed and

evaluated. It appears that the trial roulette response format improves accuracy and counters overconfidence (Goldstein et al., 2008; Goldstein and Rothschild, 2014; Haran and Moore, 2010, 2014). Johnson et al. (2010b) evaluated the trial roulette method by eliciting judgments from academic specialists about probabilities of 3-year survival with and without medicine for pulmonary hypertension patients, and concluded that the trial roulette method is feasible, has face validity, is internally valid, and has good intrarater reliability. Compared to the quantile method, the trial roulette method provides immediate visual feedback to experts, which can reduce bias, and improve reliability and validity (Clemen et al., 2000; Haran and Moore, 2014).

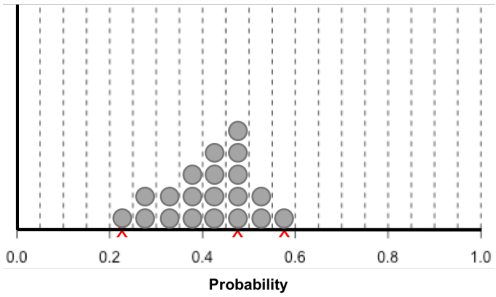


Fig. 4.2: Bins and chips method according to Johnson et al. (2010b). Experts are first asked to indicate their estimation of survival probability with an X. Subsequently, the experts are asked to indicate the lower and upper limits of their estimate using an X. Finally, experts are given 20 stickers, each representing 5% probability. Experts are asked to place the stickers in the intervals to indicate the weight of belief for their survival estimates.

The current study is the first to combine the trial roulette method to elicit distributions with insights from the literature on eliciting correlations. We will follow Johnson et al. (2010b) in an effort to evaluate our elicitation procedure. Moreover, the current study illustrates the application of the procedure, and the use of the elicited distributions in a social science application.

4.1 Evaluation of the Elicitation Procedure

In the current section we evaluate the elicitation procedure using the responses and feedback from experts who participated in an illustrative elicitation event according to the elicitation procedure. The elicitation concerned the correlation between cognitive potential (i.e., IQ) and educational performance at a specific school in the Netherlands that provides special education to youth who show severe behavioral problems. This school serves two important populations: youth with an autism spectrum disorder

(ASD), and youth with diagnoses other than ASD from the diagnostic and statistical manual of mental disorders (American Psychiatric Association, 1994). Educational performance was operationalized as the youth's didactic age equivalent divided by didactic age (DAE/DA). This measure is widely used among behavioral scientists working in Dutch education to assess academic progress relative to received months of education.

4.1.1 Material and Methods

Participants

In our illustration, the expert identification and selection were conducted at once by our key informant regarding the subject matter: WH. WH is a school psychologist who works with the population of interest, and is a member of the Dutch Association of Psychologists - section Crisis Response Network School Psychologists. WH selected six behavioral scientists working on schools for youth with severe behavioral problems in The Netherlands, who were familiar with the school and population of interest. Following Hora and Von Winterfeldt (1997), the selection was based on expertise, understanding of the problem area, and statistical understanding. All six experts were contacted by e-mail, and agreed to participate, but two of them could not attend the scheduled meeting. The attending experts were 27, 33, 40, and 46 years old females, and were working as behavioral scientists for 4, 9, 18, and 16 years respectively.

Expert judgment elicitation

The procedure to elicit judgments about correlations is a semi-structured face-to-face group interview. The semi-structured setup of the procedure implies that experts are actively invited to contribute. Furthermore, the facilitator responds to questions and elaborates explanations such that everything is clear to each of the experts, which promotes validity. Group interviews additionally improve judgment synthesis through the interaction that occurs among experts, and may diminish overconfidence (O'Hagan et al., 2006; Johnson et al., 2010a).

The procedure was developed through repeated communication with colleagues at the department of methods and statistics at Utrecht University (UU), students of the research masters methodology and statistics for the behavioral, biomedical, and social sciences, and our key informant WH. Furthermore, a pilot test was conducted with students of the UU research masters Development and Socialization in Childhood and Adolescence. Details on the development of the procedure are provided as online Supplementary Material (Part I). Based on O'Hagan et al. (2006), Johnson et al. (2010a), and Johnson et al. (2010b), the elicitation procedure consists of seven phases: (1) motivation, (2) clarification, (3) education, (4) instruction, (5) background questions, (6) elicitation of expert judgments, and (7) evaluation. Instructions for the elicitation procedure are provided in Appendix A.1. The material supporting the elicitation procedure is provided as online Supplementary Material (Part II).

The first four phases of the elicitation procedure serve to improve experts' motivation for the elicitation task, and to improve their understanding of the elicitation subject, correlations, and the elicitation procedure. These elements have been shown to improve validity of elicitation processes (Clemen et al., 2000; Johnson et al., 2010a; O'Hagan et al., 2006). Experts are asked for their knowledge on the topics of interest, and are invited to complement each other's answers. In the fourth phase (i.e., instruction), the experts are given pencils with attached erasers and are assured that they can revise their answers at any time to further reduce bias (Johnson et al., 2010a). Subsequently, in the fifth phase, the experts answer some background questions about their working experience.

In phase six, the elicitation phase, the facilitator reads the questions aloud and the experts answer the same question simultaneously. It should be stated that experts can discuss their answers together or think out loud. The first task, as a warming up, is for the experts to select the most plausible correlation value from a set of illustrated correlation categories (see Figure 4.3). The illustrated categories are based on a picture from MathIsFun.com (Pierce, 2014), which is also used in the education phase to explain the concept of correlations. Specifically, in our application the experts received the following question with Figure 4.3:

> "1. How strong do you think the relation between IQ and the ratio of didactic age equivalent with didactic age (DAE/DA) is for students at school X with an autism spectrum disorder? And for students at this school with another DSM-IV diagnosis (e.g., ADHD, ODD, attachment disorders, etc.)? Circle the best fitting correlation for both groups."

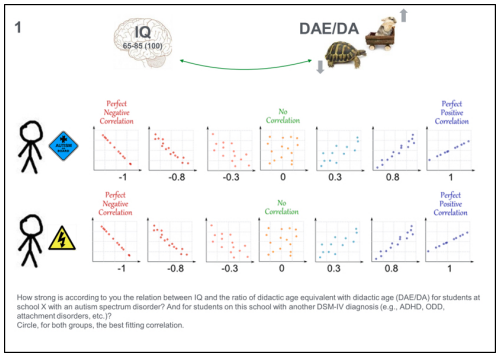


Fig. 4.3: Material for elicitation question 1: Eliciting a point estimate by selecting the best fitting correlation category for two groups.

the experts received the name of the school of interest, but for privacy reasons the name of the school is not published

The second question is the trial roulette question. As a first step, experts are asked to indicate the magnitude of the relationship of interest with a cross on a continuous scale ranging from −1 to +1 (see Figure 4.4). Specifically, the question in our application was:

> "2a. In the previous question you provided an estimate of the relation between IQ and DAE/DA for students enrolled at school X with and without autism spectrum disorder. Indicate with a cross on the A3 paper how strong you think this relation is for both groups when you can choose from all values between -1 and 1."

Subsequently, they were asked:

> "2b. Maybe you are insecure about the estimates you just provided. Indicate on the axis at the previous page also what your lower and upper limit for this estimate would be."

Finally, the experts receive 20 removable stickers (⊘ = 8 mm), each representing 5% probability, to indicate the plausibility of values between their lower and upper limit. The written instruction they receive is:

> "2c. Use the 20 stickers to indicate the weight of your expectation at every place between those limits (further instruction is provided by the facilitator)."

The facilitator explains that stickers can overlap horizontally to represent a very dense distribution. The stickers, however, cannot overlap vertically, because the height of the distribution represents probability, and each sticker represents 5% irrespective of the vertical overlap. The stickered distributions are the target of the trial roulette question, and the main output of the elicitation procedure.

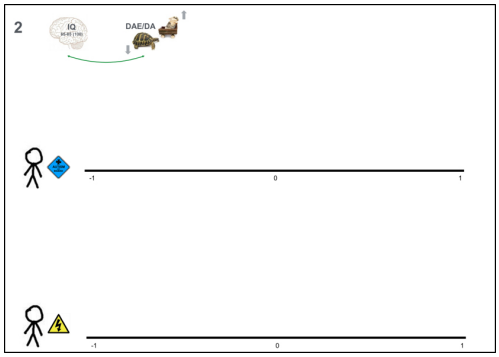


Fig. 4.4: Material for elicitation question 2: Scale ranging from -1 to +1 on which experts indicate (1) a point estimate, (2) a lower and upper limit, (3) the probability of all values by means of 20 stickers each representing 5%.

The third question is a feedback question to help the experts reflect on their trial roulette responses, and adjust their answers when necessary. The feedback question assesses concordance probability (Gokhale and Press, 1982). When we let X_i denote educational performance of student i, and Y_i cognitive potential of student i, then concordance probability inquires the probability that $Y_2 > Y_1$ given that $X_2 > X_1$. According to Clemen et al. (2000), assessing concordance probability is the second best method to elicit correlations. Specifically, the experts answered the following questions in our application:

> "3a. Imagine we select a hundred times two random students with autism spectrum disorder at school X. How often out of hundred do you think that the one with the highest cognitive potential also has the best educational performance?"

> "3b. Imagine we select a hundred times two random students with a DSM-IV diagnosis other than autism spectrum disorder at school X. How often out of hundred do you think that the one with the highest cognitive potential also has the best educational performance?"

The experts are asked to disregard previous responses in answering this question to let it function as a proper feedback question. Hence, the relation between concordance probabilities and correlations is not explained to the experts. When every expert has written down their answer, the facilitator asks the experts for their values and translates the values into correlations using:

$$r = \sin\left(0.5[\frac{2\pi x}{100} - \pi]\right), \tag{4.1}$$

where r is the correlation, and x is the frequency as provided by the expert. The experts are asked to review and adjust their stickered distributions considering their answers to the concordance probability question. When the experts are satisfied with their judgment distributions, they can continue to the evaluation phase of the elicitation procedure. The questions asked in this phase are specified in the next section.

Elicitation event

The elicitation event took place at a school for youth with behavioral problems where all experts had a meeting scheduled that day. Before the start of the elicitation, all experts gave permission to audio-record the elicitation. The duration of the elicitation event was 40 minutes.

Assessment of measurement properties

When expert judgments are elicited, validity indicates that the distributions accurately reflect the uncertain knowledge of the experts (Van Lenthe, 1993). In the elicitation

procedure, validity is therefore assessed with questions about the elicitation procedure to the experts. More specifically, in our application face validity was assessed with the following question:

> "To what degree do you feel that your expert-knowledge about the relation between cognitive potential and educational performance was assessed accurately?"
> Not at all / Not really / Neutral / A little bit / Completely

Feasibility is assessed by two statements. The first statement is:

> "I thought the questions with their explanations were clear."
> Not at all / Not really / Neutral / A little bit / Completely

The second statement is:

> "I thought the questions were easy to answer / conduct."
> Not at all / Not really / Neutral / A little bit / Completely

After each question and statement space is provided to add an explanation. The mean scores over experts were calculated for the two statements, and the average was taken as a final estimate of feasibility. Additionally, the participants answer an open follow-up question:

> "Which question did you find the least clear, and why?"

Furthermore, the correlation among individual experts' responses on the trial roulette question and the concordance probability feedback question was computed to assess convergent validity between questions within the procedure. Additionally, the absolute differences between experts' responses on the trial roulette question and the concordance probability feedback question were calculated as another measure of convergent validity. Subsequently, the coherence among experts with respect to the same question was evaluated as an indication of validity, since we expect experts do agree to a certain extent. Finally, a retest was conducted to assess test-retest reliability. All calculations were conducted in R (R Core Team, 2015). Relevant data and R-code are provided as online Supplementary Material (Part III).

4.1.2 Results

The elicitation event proceeded as planned. The experts discussed their views on the population and measures in the clarification phase, and indicated that they understood everything explained in the education phase. During the first question to elicit correlations, the experts discussed the direction of the correlation, and they mentioned that their preferred correlation was not amongst the answer categories. Additionally, they discussed differences among IQ tests. During the second and third question, the experts mainly discussed the task, but not their judgments. One expert varied the vertical distance between stickers substantially, which was noted by the facilitator and adjusted by the expert.

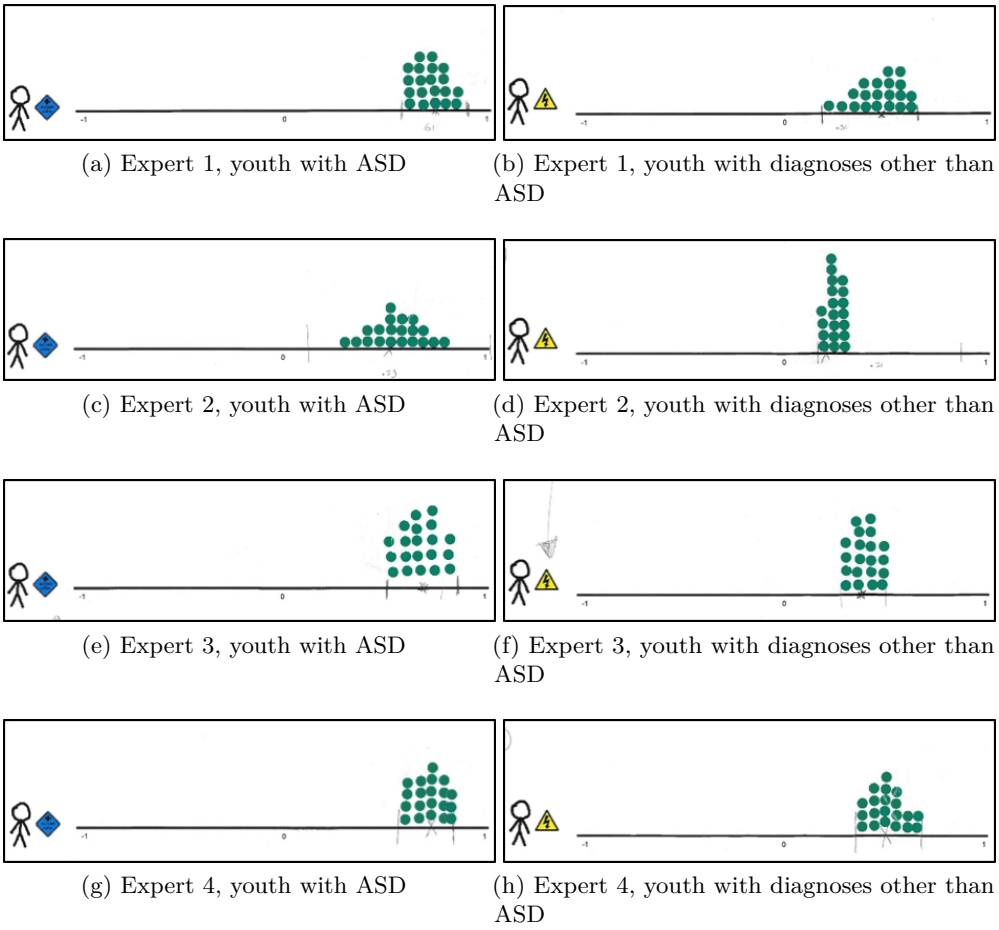
Figure 4.5 shows the elicited distributions for all experts (rows) by evaluated target population (columns), and Table 4.1 shows the experts' point estimates. The distributions depicted in Figure 4.5a, 4.5c, 4.5e, and 4.5g show that for youth with ASD the correlation between cognitive potential and educational performance is between .29 and .79 according to expert 2, while the other experts expect the correlation to be .5 or higher, up to .86. For youth with diagnoses other than ASD (Figure 4.5b, 4.5d, 4.5f, and 4.5h), expert 2 is most specific and expects the correlation to be between 0.16 and 0.31. The other experts are somewhat more uncertain, and expect somewhat higher correlations, but all expect that the correlation for youth with ASD is likely larger than that for youth with other DSM-IV diagnoses.

Table 4.1: Elicited point estimates and their absolute differences for the correlation derived from question 2a, and question 3 on concordance probability

	r ASD (Q2a)	r ASD (Q3a)	Δ	r no ASD (Q2a)	r no ASD (Q3b)	Δ
Expert 1	.725	.612	.112	.457	.249	.226
Expert 2	.525	.588	.063	.200	.309	.109
Expert 3	.675	.707	.032	.375	.309	.066
Expert 4	.725	.809	.084	.500	.588	.088

Q2a refers to Question 2a where the expert is asked to provide a point estimate for the correlation.
Q3a refers to Question 3a which requires the expert to provide a frequency for the concordance probability for youth with ASD.
Δ refers to the absolute difference between the two previous columns.
Q3b refers to Question 3b which requires the expert to provide a frequency for the concordance probability for youth with diagnoses other than ASD.

The raw data was digitalized after the procedure described in Appendix A.2. Figure 4.6 and 4.7 display the digitalized distributions of the experts in four ways for youth with ASD and youth with diagnoses other than ASD, respectively. Figures 4.6a and 4.7a show the distributions as histograms, which can be directly used as priors in a Bayesian analysis (Albert, 2009), but this is not a straightforward option in current software. Another way to process the results is as distributions with a known form; parametric distributions (see Figure 4.6b and 4.7b). Parametric distributions can be derived from histogram distributions by means of the Sheffield Elicitation Framework R file (SHELF; Oakley and O'Hagan 2010). Specific code, and the equations for the parametric priors are provided in Appendix A.3. Parametric distributions can be used directly as priors in a Bayesian analysis in most Bayesian software. The information provided by the histograms and parametric distributions is similar to that of the raw data as described in the previous paragraph.



(a) Expert 1, youth with ASD

(b) Expert 1, youth with diagnoses other than ASD

(c) Expert 2, youth with ASD

(d) Expert 2, youth with diagnoses other than ASD

(e) Expert 3, youth with ASD

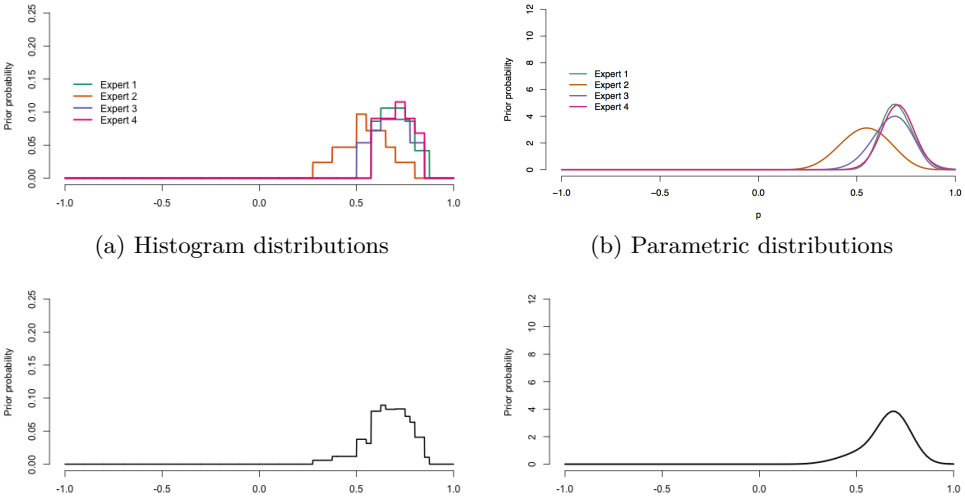
(f) Expert 3, youth with diagnoses other than ASD

(g) Expert 4, youth with ASD

(h) Expert 4, youth with diagnoses other than ASD

Fig. 4.5: Trial roulette responses for the correlation between cognitive potential and educational performance for youth with ASD and youth with diagnoses other than ASD enrolled in special education for youth with severe behavioral problems.

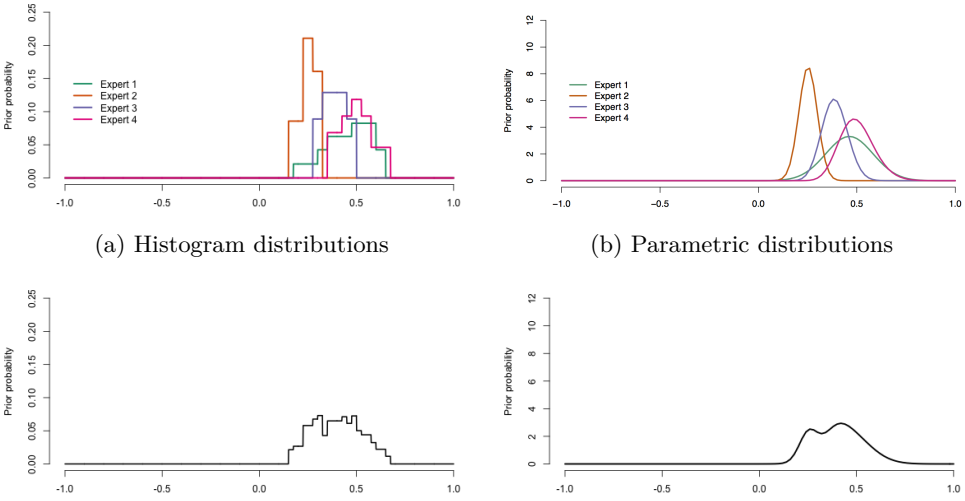
The histogram and parametric distributions of the separate experts can also be pooled to obtain an idea of the judgments of the experts as a group. One method to aggregate the distributions is linear pooling (Genest and Zidek, 1986). Linear pooling is a method in which the (weighted) average distribution is calculated. The determination of weights received considerable attention in the literature. For example, experts can be assigned equal weights, experts can be ranked, experts can rank themselves and weights can be attributed proportionally to this ranking, or a performance based method such as the the Classical Model (Cooke, 1991) can be applied (Winkler, 1968). The



(a) Histogram distributions (b) Parametric distributions

(c) Pool of histogram distributions (d) Pool of parametric distributions

Fig. 4.6: Digitalized expert judgments for youth with ASD.



(a) Histogram distributions (b) Parametric distributions

(c) Pool of histogram distributions (d) Pool of parametric distributions

Fig. 4.7: Digitalized expert judgments for youth with diagnoses other than ASD.

Classical Model determines weights based on a score for calibration and information. This method requires relevant seed variables for which the truth is or becomes known. In the current study we wanted the prior to reflect the current view of the experts as a group, hence we chose equal weights. The pooled histogram distributions obtained with equal weights are shown in Figure 4.6c and 4.7c, and further explained in Appendix A.4. Figure 4.6d and 4.7d show the pooled parametric distributions. More details on the linear pool of parametric distributions are provided in Appendix A.5. For the population with ASD the mode for the correlation is around .67, and the 95% interval of values that the experts put most weight on ranges from about .41 to .86. The population with diagnoses other than ASD does not have one clear mode, but the 95% interval ranges from about .18 to .64 in both the histogram (Figure 4.7c) and parametric pooled distribution (Figure 4.7d).

Validity

The four experts rated face validity with 4, 2, 4, and 4 respectively on a scale from 1 to 5. The expert that provided the lowest score wrote in the open space after the question about the accurateness of the assessment: "More engaged with the statistics → are your own answers reliable? It has to be correct". The expert's comment was interpreted as indicating that transforming her ideas into proper responses was more difficult than forming judgments, which is not problematic as long as she was satisfied with the final result. The average face validity score of 3.5 was interpreted as satisfactory.

The experts provided scores of 4, 5, and 5 for clarity, and 4, 4, 4, and 5 for ease of of the questions. The average score for feasibility was thus 4.46. The expert that provided the 4 for clarity added that once she had thought about the questions, they were clear. One expert did not provide a score for clarity and added that the verbal explanations were absolutely necessary for her. The feasibility score was interpreted as excellent, because verbal explanations are part of the procedure. Two experts indicated which question they found least clear. One expert wrote that question 1 was the least clear, and explained that this question contained a mistake. Indeed, the question referred to DA/DAE instead of DAE/DA, but this was clarified when the question was addressed, so it will not have affected the validity of the responses. The other expert wrote that question 2 was the least clear, but did not explain her response.

Convergent validity between questions within our procedure was first evaluated by correlating the experts' trial roulette point estimates (Table 4.1, column 1 & 4), and their answers to the concordance probability question converted to a correlation by means of Equation 1 (Table 4.1, column 2 & 5). Note that the experts were asked to reconsider their probability distribution after obtaining a correlation value for their concordance probability response, but did not adjust their initial point estimate. With respect to adolescents with ASD, the correlation between the responses to both questions was .59, (SE = .57). The Bayes factor quantifying the relative evidence for a positive correlation versus a correlation of zero as calculated by JASP 0.8.0.0 (JASP Team, 2016) with default priors was 1.2. With respect to adolescents with other DSM-IV diagnoses, the correlation was .42 (SE = 0.64), and the Bayes factor was 0.9.

The point estimates are an indication of sufficient convergent validity. However, the standard errors show that with four participants the estimates must be interpreted with caution. Additionally, the Bayes factors suggest that there is more evidence for a positive correlation for the first population, but more evidence for a correlation of zero for the second population.

Correlations can be perfect when a bias is systematic, therefore, the absolute difference between the two point estimates may be an even more important indication of convergent validity. The differences between estimates from the trial roulette and concordance probability question are provided in column 3 and 6 of Table 4.1. Over the two populations, the difference was on average .10 (.07 and .12 for the population with and without ASD respectively), which we consider a small difference, and thus a positive indication of convergent validity.

Since the trial roulette method is implemented in the procedure because of the distributions it provides, we also comment on convergent validity between the concordance probability results and the raw distributions (Figure 4.5). We note that all point estimates given in Table 4.1 fall within the distributions specified in Figure 4.5, which means that the point estimates provided in the concordance probability questions are also among the plausible values in the accompanying trial roulette response. These matching responses are a positive indication of convergent validity, but note that participants were allowed to adjust their distributions after receiving feedback from the concordance probability question.

The coherence between the judgment distributions of different experts was taken as a measure of validity. Figure 4.6a shows that for the population with ASD, the expert judgments clearly cluster and overlap supporting the validity of the procedure. Figure 4.7a shows that for the population with diagnoses other than ASD the judgments also cluster, but the judgments of expert 2 and expert 4 do not overlap. Since the judgments of expert 2 and expert 4 both overlap with expert 1 and expert 3, it was considered an indication of sufficient validity. To further improve the coherence between expert judgments, the facilitator could encourage the experts to discuss their answers and distributions. The facilitator could, for example ask an expert:"Can you tell me about your distribution and explain the decisions that you have made?".

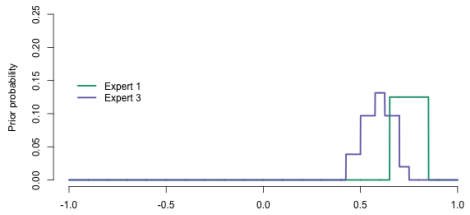
Reliability

To assess test-retest reliability, the experts were sent the same questionnaire by mail 4.5 months after the original elicitation event. Three out of the four experts were able to respond within four weeks. The responses, however, were different than expected: Questions were skipped (expert 2), and the experts (expert 1 and 3) that provided an answer to the concordance probability question for youth with diagnoses other than ASD provided values that correspond to negative correlations, which was inconsistent with their other responses within the retest and original elicitation.

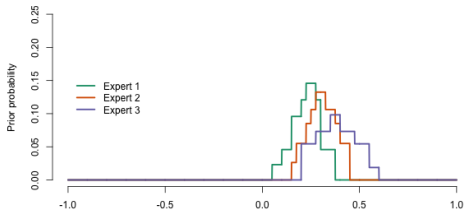
Following Johnson et al. (2010b) the ICC (2,k) of Shrout and Fleiss (1979) was calculated over the point estimates of the three responding experts as a measure of intrarater reliability. The ICC was 0.22 [-0.27, 1.00] with respect to youth with

ASD. An ICC value of 0.6 would be moderate. For youth with diagnoses other than ASD, the ICC did not provide sensible values: -1.67 [-4.35, 1.00], because the residual variance was larger than the variance between occasions. Thus, intrarater reliability was insufficient with respect to the point estimates.

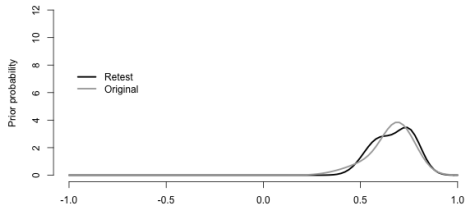
Distributions for youth with ASD were only provided by expert 1 and 3 in the retest (see Figure 4.8a). For youth with other diagnoses, expert 2 stickered a shape instead of a histogram. Nevertheless, we were able to digitalize it in the form of a histogram prior, giving Figure 4.8b. The pooled retest and original distributions are provided in Figure 4.8c and 4.8d. Despite the inconsistencies in the concordance probability and correlation point estimates, the trial roulette distributions in the retest were similar to the distributions in the original test.



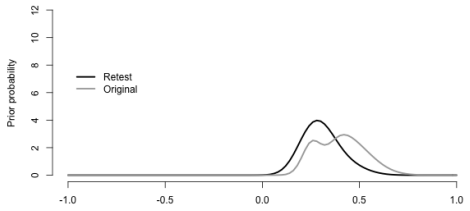
(a) Youth with ASD



(b) Youth with diagnoses other than ASD



(c) Pool for youth with ASD



(d) Pool for youth with diagnoses other than ASD

Fig. 4.8: Digitalized expert judgments retest.

In sum, conventional and custom measures of face-validity, feasibility, convergent validity, and coherence provided positive indications for the validity of the elicitation procedure. The results of the retest were less positive and raise a number of possible interpretations: the poor results for the point estimate reliability and inconsistent concordance probabilities may show that test-retest reliability is low, or that a face-to-face group process is important for consistent responses. The experts may have had difficulties to make time and concentrate on the task in their own environments, making them struggle to conduct tasks that they managed to do in the group setting.

The trial-roulette distributions showed better test-retest reliability, although one of the experts did not put the stickers according to the instructions. In the next section, the practical use of the elicitation is illustrated with an empirical application.

4.2 Use of the Elicitation Procedure

Following Figure 4.1, the current section provides a full description of the empirical application to illustrate the practical use of the elicitation procedure in a model with the correlation as its key parameter.

4.2.1 Step 1. Question

The objective in this application was to update the current knowledge of behavioral scientists working in special education for youth with severe behavioral problems about the correlation between cognitive potential and educational performance for two populations at a specific school in the Netherlands. The populations of interest were (1) youth enrolled in special education, because of severe behavioral problems, who have autism spectrum disorder (ASD), and (2) youth enrolled in this type of special education without ASD but with other DSM-IV diagnoses. Examples of DSM-IV diagnoses that youth in the second population have are oppositional defiant disorder (ODD), attention deficit hyperactivity disorder (ADHD), and attachment problems.

Youth enrolled in special education for reasons of severe behavioral problems are a population that is difficult to recruit, because they are considered vulnerable and are subjected to tests more often than most of them desire. To let these youths participate, informative consent is required from the adolescents themselves as well as a parent or legal guardian in case the adolescent is younger than 16. The files that contained the necessary information for our research, contained personal, and often sensitive information, which increases reluctance to participate. As a result, we expected to gather only a small amount of data. On itself, limited data as obtained in the current application can provide little information, but in combination with expert judgments, it can increase the confidence in current expert views, or indicate that adjustments of these views might be relevant, which can also be an impulse for new research.

Ethical approval for the elicitation procedure, and data collection was given by the Ethics Committee of the Faculty of Social and Behavioral Sciences Utrecht (FETC). Informative consent was obtained from the adolescents. When adolescents were younger than 16 years, informative consent was also obtained from a parent or legal guardian.

Cognitive potential was operationalized as intelligence quotient (IQ) measured with the Wechsler Intelligence Scale for Children (WISC-III; Wechsler 1991). Educational performance was operationalized as the youth's didactic age equivalent divided by didactic age (DAE/DA).

4.2.2 Step 2: Elicit Expertise

The expert sample is described in Section 4.1.1. The elicitation procedure is described in Section 4.1.1 and Appendix A.1.

4.2.3 Step 3: Construct Priors

In the current section, we explain how we constructed priors for all parameters in the bivariate normal distribution: the correlation, the means of DAE/DA and IQ, and the standard deviations of these variables.

The prior for our key parameter, the correlation, was derived from the experts' trial roulette responses for both populations (see Figure 4.5 for the raw judgment distributions, and Figure 4.6 and 4.7 for the digitalized judgment distributions). Since our research goal was to update current expertise, and not expertise specifically related to one expert, we preferred a pooled distribution as a prior. As we show in Appendix Section A.6.4, combining a pooled prior distribution with data in one analysis gives a posterior result equal to pooling posteriors of analyses in which each expert's judgment distribution was combined with the data separately. Since the latter approach is more straightforward in software currently available, this approach was adopted in the current study. While the pool of histogram distributions (Figure 4.6c, and Figure 4.7c) seemed very similar to that of parametric distributions (Figure 4.6d, and Figure 4.7d), we preferred the pool of parametric distributions because parametric distributions are also more straightforward to deal with in current software, which seems relevant for future users of the procedure.

Priors were also composed for the means (i.e., μ) and standard deviations (i.e., σ) of IQ and DAE/DA. The rational for the prior of μ_{IQ}, $p(\mu_{\text{IQ}})$, was based on literature. Expert judgments could have been elicited for the other parameters in the model too, but our experts lacked the time for further elicitation practices. Therefore, we made use of the literature to specify sensible prior distributions for these parameters. Youth who are enrolled in special education because of severe behavioral problems score well below average on IQ. The WISC-III uses the following IQ-score classifications: intellectually deficient, borderline, low average, average, high average, superior, and very superior (Weiss et al., 2006). The borderline class was considered most appropriate for our population. The accompanying IQ scores for this class are 70-79. The rounded class middle of 75.0 was considered a good estimate for the average IQ in our population. A variance of 400.0 ($SD = 20.0$) was chosen to construct a prior distribution with its first quartile at 61.51 and third quartile at 88.49. In addition, the distribution was truncated at the values 45.0 and 145.0, since these values constitute the range of the WISC-III. Thus the equation for the prior was as follows: $p(\mu_{\text{IQ}}) \sim N(75.0, 400.0)I_{\mu_{\text{IQ}} \in [45,145]}$.

The rationale for $p(\sigma_{\text{IQ}})$ was that the standard deviation of IQ is by definition 15.0 in the population (Prifitera and Saklofske, 1998). A common prior for standard deviations is the gamma prior. The shape and rate parameter of the gamma distribution for the standard deviation of IQ were specified such that the first and third quartile of

the distribution were 9.57 and 19.28 respectively ($M = 15.09$). Thus, the equation for the prior was as follows: $p(\sigma_{\text{IQ}}) \sim \Gamma(2.0, \frac{1}{7.5})$.

With respect to $p_{\mu_{\text{DAE/DA}}}$ we know that youth following special education for reasons of severe behavioral problems generally lag behind, and thus have a DAE/DA below 1.0. As a rough estimate for the average DAE/DA 0.75 was chosen. The variance of the mean was specified to be 0.5. With this specification, the first and third quartile of the prior distribution were 0.27, and 1.23 respectively. The distribution was truncated at 0.0 and 1.5, because more extreme values are naturally impossible to constitute the average for the population of interest. Thus, the equation for the prior was as follows: $p(\mu_{\text{DAE/DA}}) \sim N(0.75, 0.50)I_{\mu_{\text{DAE/DA}} \in [0.0,1.5]}$.

To our knowledge, no literature exists about $\sigma_{\text{DAE/DA}}$. However, on a scale of 0.0 to 1.5, we considered a standard deviation of 0.36 most likely. A standard deviation of 0.36, namely, would create a 95% confidence interval ranging from 0.04 to 1.46, which constitutes 95% of a normal distribution that ranges from 0.0 to 1.5 with a mean value of 0.75. The shape and rate parameters for the gamma distribution were specified such that the first and third quartile were 0.17, and 0.49 respectively ($M = 0.36$). Thus, the equation for the prior was as follows: $p(\sigma_{\text{DAE/DA}}) \sim \Gamma(2.0, 5.5)$.

4.2.4 Step 4: Collect New Data

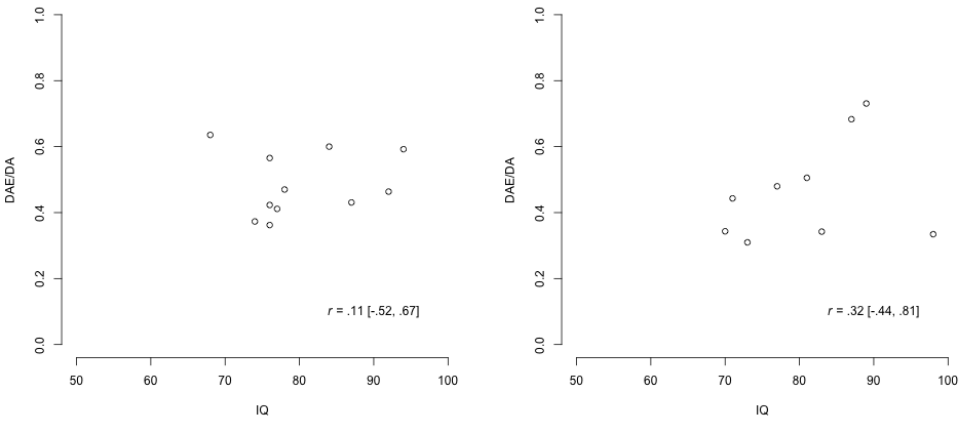
We obtained informed consent for 28 adolescents enrolled at a Dutch secondary school for youth with severe behavioral problems to collect information on the research variables of interest from the personal records of the adolescents. For 20 adolescents, the records contained the required data on DSM-IV diagnoses, DAE, DA, and IQ were retrieved from participants' school records. DAE was separately reported for technical reading, reading comprehension, spelling, arithmetic, and vocabulary. An average DAE-score was calculated when scores for at least three of the subjects were available, otherwise, the DAE was regarded missing. When multiple IQ-scores were available, the most recent WISC-III score was included.

Eleven out of the 20 adolescents for which sufficient data was present (10 male, 90.9%) belonged to the sample with ASD, and nine (6 male, 66.7%) belonged to the sample with diagnoses other than ASD. The data for DAE/DA and IQ are plotted in Figure 4.9. As expected, the amount of data was very limited, and would provide little information on the correlations of interest. However, in combination with the expert judgments, it could increase confidence in current expert views or indicate that adjustments of these views are relevant.

4.2.5 Step 5: Update

Analysis

In a Bayesian analysis, the prior distribution is multiplied with the (density of the) data, resulting in a posterior distribution. We conducted our analyses with the software JAGS (Plummer, 2013) via the package rjags (Plummer, 2015) in R (R Core Team, 2015).



(a) Youth with ASD (b) Youth with diagnoses other than ASD

Fig. 4.9: Scatter plots of the data for DAE/DA and IQ, where r indicates the correlation in the data.

In Appendix A.6, we specify the elements of the analyses, and relevant information to properly report a Bayesian analyses (Depaoli and Van de Schoot, 2017). Annotated R-code and anonymized data to replicate the results is provided as online supplementary material (Part IV).

Results

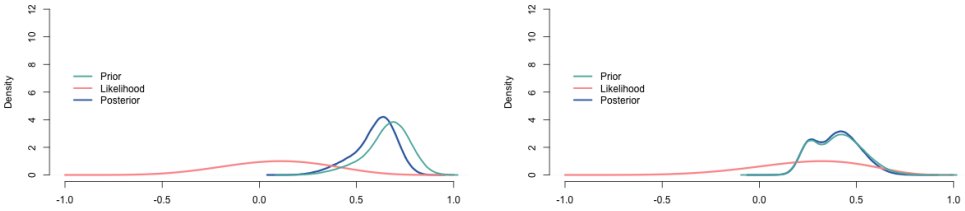
For the population with ASD, Table 4.2 summarizes the judgments of the experts, the correlation in the data, and the resulting posteriors. The means of the posterior distributions are all lower than those of the prior distributions as an effect of the low correlation in the data. Another result is that the posterior distributions are more specific than the accompanying priors and the correlation in the data by themselves are. To finish the analysis, we combined the separate posterior distributions for the correlation and constructed the pooled posterior distribution. The pooled posterior distribution for the correlation is displayed in Figure 4.10a, and summarized in the last column of Table 4.2. Figure 4.10a also depicts the aggregated prior, and the (relative profile) likelihood (Bertolino and Racugno, 1992) of the correlation in the data.

For the population with diagnoses other than ASD, Table 4.3 summarizes the judgments of the experts, the correlation in the data, and the resulting posteriors. The means of the posterior distributions are similar to those of the prior distributions, because the correlation in the data is of a similar size as well. Again, the posterior

Table 4.2: Elements of the updating procedure: prior per expert, pooled prior, correlation in the data, posterior per expert, and the pooled posterior for the correlation for the population with ASD

	M prior r [95% HPD]	M pooled prior r [95% HPD]	M data r [95% CI]	M posterior r [95% HPD]	M pooled posterior r [95% HPD]
Expert 1	.71 [.55, .87]	.66 [.40, .87]	.11 [-.52, .67]	.65 [.52, .78]	.59 [.35, .79]
Expert 2	.54 [.31, .78]			.48 [.26, .68]	
Expert 3	.68 [.49, .86]			.60 [.42, .78]	
Expert 4	.71 [.55, .87]			.65 [.51, .79]	

Note. HPD refers to highest probability density. CI refers to confidence interval.



(a) Youth with ASD (b) Youth with diagnoses other than ASD

Fig. 4.10: Visualization of the prior, the relative profile likelihood, and the posterior distribution for the correlation.

distributions are more specific than the accompanying priors and the correlation in the data by themselves are. As for the population with ASD, we finished the analysis by constructing the pooled posterior distribution. The pooled posterior distribution for the correlation is displayed in Figure 4.10b, and summarized in the last column of Table 4.3. Figure 4.10b also depicts the aggregated prior, and the (relative profile) likelihood (Bertolino and Racugno, 1992) of the correlation in the data.

We investigated the impact of the priors for the standard deviations by means of a sensitivity analysis as advised by Depaoli and Van de Schoot (2017) in their checklist for transparent and replicable Bayesian research. The alternative prior that we used was $\Gamma(0.01, 0.01)$, which is a regular prior for standard deviations. The results show that the posterior distribution is hardly affected by our choice of priors. For the standard deviation of DAE/DA in the population with ASD, the means of the posterior distribution are 0.13 or 0.14 for the regular and informative priors respectively. For

Table 4.3: Elements of the updating procedure: prior per expert, pooled prior, correlation in the data, posterior per expert, and the pooled posterior for the correlation for the population with diagnoses other than ASD

	M prior r [95% HPD]	M pooled prior r [95% HPD]	M data r [95% CI]	M posterior r [95% HPD]	M pooled posterior r [95% HPD]
Expert 1	.46 [.22, .70]	.40 [.18, .64]	.32 [-.44, .81]	.44 [.23, .66]	.39 [.18, .61]
Expert 2	.46 [.22, .70]			.25 [.16, .34]	
Expert 3	.39 [.26, .52]			.39 [.26, .51]	
Expert 4	.50 [.34, .68]			.49 [.34, .65]	

Note. HPD refers to highest probability density. CI refers to confidence interval.

the population with diagnoses other than ASS the means of the posterior distributions are 0.18, and 0.19. For the standard deviation of IQ, the means of the posteriors are 10.04 and 10.50 for the regular and informative prior respectively. For the population with diagnoses other than ASS, the means of the posterior distributions are 9.91, and 10.38 for the regular and informative prior respectively.

4.2.6 Step 6: Evaluate

The pooled posterior distributions reflect the result of updating the judgments of experts with data. The posterior distributions for both populations are compromises between the prior judgments of the experts and the information in the data. The posterior distributions have smaller 95% intervals than either the pooled prior or the data alone, because our confidence increased by combining the two sources of information. Interesting to note is that the data only slightly affected the posterior distributions for both populations. This small impact is caused by the limited amount of information that can be derived from 11 or 9 data points. The relatively flat and wide likelihood distributions (Figure 4.10) illustrate this nicely.

According to the updated state of knowledge, the correlation between cognitive potential and educational performance is most likely large for youth with ASD who are enrolled in special education because of severe behavioral problems. By updating the expert judgments with new data, the judgment about the correlation has been slightly modified downwards. This modification raises the question whether additional data would again have such an effect. A new research cycle may be started based on this question. With respect to youth with diagnoses other than ASD, updating the expert judgments with new data slightly modified, but mainly reinforced current expert views of a medium correlation between cognitive potential and educational performance. New data and new experts may further update this adjusted judgment.

Following the research cycle, the school in question gained insight into the views of school psychologists with respect to the relation between educational performance

and cognitive potential for two of the populations that visit the school, and the fusion of these views with local data. A new research cycle may be started to further update the current state of knowledge.

4.3 Discussion

The purpose of the current paper was to evaluate and apply a procedure to elicit judgments for correlation priors. The results of this procedure using the trial roulette method are promising. Measures of face-validity, feasibility, convergent validity, coherence, and intrarater reliability showed positive results. Furthermore, the results of the procedure were useful as prior information in a Bayesian analysis.

The proposed elicitation procedure can be used to elicit experts' prior judgments about Pearson's product moment correlations for bivariate models. For models with more variables, conditional correlations need to be elicited to retain a positive definite correlation matrix. Further research is required to see if the trial roulette method is also suited to elicit the conditional correlations. The elicitation of conditional correlations increases in complexity as the size of the correlation matrix increases. Werner et al. (2017) wrote a review on expert judgment for dependence that offers guidance on making choices about summaries of expert knowledge for multivariate distributions.

Several digital trial roulette elicitation tools have been developed. For example, SPIES (Haran and Moore, 2014), and the MATCH Uncertainty Elicitation Tool (Morris et al., 2014). Advantages of these elicitation tools are that they can be easily distributed, and there is no need to digitalize the elicited responses anymore. On the other hand, the digital mode is less suitable for discussion among experts, and providing additional explanation when necessary. Since correlations are considered more complex than probabilities, an interactive (face-to-face) education phase may be more important in this context. Given that experts in our study skipped questions and ignored instructions outside the group setting, we expect that the suitability of digital elicitation differs between populations of experts.

People tend to vary their responses depending on the specific "anchors" (i.e., fixed values) they are provided with (O'Hagan et al., 2006). To avoid too much influence on the judgment process from "random" values, we chose to provide only three tick labels at meaningful points (-1, 0, 1) along the scale of the trial roulette question. A potential issue raised by one of the reviewers, however, was that not providing more tick labels may have lowered the validity of the trial roulette question, since experts may have been unable to pinpoint specific values along the line. Further research is required to investigate whether it is important for valid responses that experts know to what correlation value the points along the axis correspond. If it is important for experts to have more tick labels, it should be investigated how many tick labels are required, and whether they should be evenly distributed along the scale, or be placed at meaningful values like Cohen's (1988) indications of small, medium, and large correlations. A potential increase in validity by placing tick labels should be balanced with the loss in validity that could be induced by anchoring.

The evaluation of the elicitation procedure and the illustrative application have limitations. Most importantly, only four experts participated in the final elicitation procedure. Four experts can be sufficient, but a panel of about eight is recommended (Cooke and Goossens, 1999). When more experts are involved, it is easier to recognize the general opinion and the final result is less sensitive to the misjudgment of one expert. Furthermore, the identification and selection of experts generally is a process with multiple stages in which potential experts are asked to identify other experts until no new names appear. Subsequently, experts are selected based on relevant criteria. In some cases a panel may be installed to select experts based on their curriculum vitae (Cooke and Goossens, 1999). In the illustrative application of the elicitation procedure, one key informant identified and selected experts, which may limit the diversity of the expert's judgments.

Because validity is the accurate representation of experts' judgments in our research context, and our research data was not necessarily unbiased, we did not validate the accuracy of the expert judgments against data. Consequently, we cannot rule out that all experts were wrong about the truth in the population. When finding the truth about the correlation in the population is the main goal, researchers need sufficient unbiased data, or a seed variable that can indicate the accuracy of the experts' judgments (Cooke, 1991).

Considering the distributions of the experts, one might suspect overoptimism (i.e., expecting the effect to be larger than it is in reality) and overconfidence (i.e., specifying too narrow intervals) to play a role. Goldstein and Rothschild (2014), however, showed that even laymen can properly retrieve underlying population distributions about frequencies. Overoptimism can also be reduced by pooling over experts (Johnson et al., 2010a) as we did in the current study. Additionally, the feedback by the concordance probability question can help experts to detect potential overoptimism. Overconfidence may very well be an issue in the experts' judgments. SPIES has shown to reduce overconfidence compared to directly asking for intervals or fractiles, but even in this method 90% intervals cover the truth in only 73.8% of the cases (Haran and Moore, 2010). It may be worthwhile to introduce extra variance in prior distributions based on expert elicitation before updating it with data when trying to retrieve the correlation in the population.

For future use of the elicitation procedure, naturally, the variables and accompanying illustrations should be adjusted to the research questions at hand. Additionally, we would advise to ask experts to reflect on their judgments. Such a reflection creates an additional feedback moment and encourages experts to discuss their judgments, which further promotes judgment synthesis. Directions to facilitate a group discussion on expert judgments have been provided recently in SHELF 3.0 (Oakley and O'Hagan, 2016). Finally, we did not deviate from the way Johnson et al. (2010b) asks the experts about upper and lower limits. Consequently, like Johnson et al. (2010b) we are not certain whether the experts interpreted the limits of their plausible estimate as a 90%, 95%, 100%, or another confidence interval. Oakley and O'Hagan (2016) provide an instructional slideshow to explain the meaning of plausible limits to experts that can be used in future applications.

With the elicitation procedure, users can progress from having no expert judgments about the correlation at all, to distributions of probable values according to experts, which can be further updated with new data. When the expert judgments and data are alike, the updated distribution shows that experts can increase their confidence. When the expert judgments and data are more dissimilar, the expert views can be adjusted when both sources of information seem trustworthy, but it can also be an important impulse for further research. Thus, combining expert judgments with data either leads to more confident conclusions, or results in new research questions which can be further investigated according to the research cycle.

Acknowledgments

We would like to give special thanks to Julie Loth, Karen Ploemacher, Angelique van den Kerkhoff, and Roos Geerders for sharing their expertise, Horizon Youth Care and Education for officially adopting the project, the DaSCA class of 2016 for pilot testing our procedure, Maria Bolsinova for her insightful feedback, Rod Pierce of MathIsFun.com for giving permission to use the image on correlations that is also included in the final procedure, and the reviewers for their significant input.

A

Appendices

A.1 Elicitation Procedure Instructions

This appendix contains instructions for the elicitation procedure. The supporting presentation and questionnaire material is provided as online supplementary material (Part II).

A.1.1 Instructions

Terms & conditions: Ask for informed consent. Ask whether session can be recorded.

1. Motivation: Supported by slides, explain the goal of the elicitation, explain why the experts are important, explain that the process will help formalize their expertise into expectations about the correlations, explain that it is natural to be uncertain, explain that uncertainty can and should be expressed in answers as well, explain that questions can be asked at any time.
2. Clarification: Supported by slides, discuss central concepts like the research population, and the variables for which the correlation is elicited. Ask the experts: How would you describe <population of interest>? Do you ever run into <variable of interest>? Do you have an idea / could you explain what <variable of interest> stands for?
3. Education: Discuss and explain correlations with Figure 4.3.
4. Instruction: Repeat that questions can be asked at any time, answers can be revised, and questions should be answered by experts at the same time so that questions can be discussed while answering them. Provide pencils with attached erasers.
5. Background Questions: Provide the questionnaire to the experts, and start the questionnaire. Wait until everyone has finished: ensure that the experts do not continue until everyone has finished the Background Questions.
6. Elicitation: Continue with the main questions, and ensure that questions are answered carefully, and simultaneously. Read question 1 of the elicitation questionnaire aloud and verify that the question is clear to everyone. Explain that it is an introductory question: later on, every value can be chosen instead of categories. Wait until everyone has finished.

Read question 2a of the elicitation questionnaire aloud and verify that the question is clear to everyone. Wait until everyone has finished.
Read question 2b of the elicitation questionnaire aloud and verify that the question is clear to everyone. Wait until everyone has finished.
Read question 2c of the elicitation questionnaire aloud, and provide 20 stickers (for every distribution that is to be specified). Explain that they receive 20 stickers, each representing 5% that reflect probability: They should think for how probable it is that the correlation between <variable 1> and <variable 2> has of that specific value and attach stickers accordingly. Ask whether the experts are content with the distributions that they have specified. If not, they can adjust. Wait until everyone has finished.
Read question 3 of the elicitation questionnaire aloud and verify that the question is clear to everyone. Explain that question 3 is a new question, which they should think about independent of previous answers. Encourage the experts to forget everything they answered so far while responding to this question. Calculate the correlation matching the indicated concordance probability, and give this value as feedback to the expert. Ask the expert to re-evaluate the distribution in question 2 with this information. Ask whether the experts are satisfied with their answers, and continue when everyone is.

7. Evaluation: Provide experts with time and privacy to answer evaluation questions.

A.2 Digitalizing Expert Judgment Distributions to Create Histogram Priors

In this appendix the digitalization of the expert judgment distributions is explained in detail. The result of the digitalization are histogram distributions. The histogram distributions can be used as priors in Bayesian analyses, but also serve as input to obtain parametric priors (see Appendix A.3).

The digitalization of an expert's judgment distribution for the correlation consists of the following three steps:

Step 1: Divide the range of possible correlation values from -1 to $+1$ into $b = 1, ..., f$ intervals (i.e., bars) with an equal width.

Step 2: For $b = 1, ..., f$, set $p_b = 0$, where p_b denotes the probability mass assigned to a bar.

Step 3: Do for $l =, ..., k$,
If $b \in [\mathrm{LL}_l, \mathrm{UL}_l]$, then $p_b = p_b + \mathrm{TP}_l / N_l$,
where l is a layer of stickers, k the number of layers, LL_l is the first bar within layer l, UL_l is the last bar within layer l, TP_l is the total probability in layer l (number of stickers in l times .05) and N_l the number of bars in the interval $[\mathrm{LL}_l, \mathrm{UL}_l]$.

Figure A.2.1 illustrates Step 1: the application of bars to the expert judgment distribution. Here, $f = 80$, but only 40 bars are projected on the distribution itself to promote clarity, the 80 bars are displayed below the x-axis.

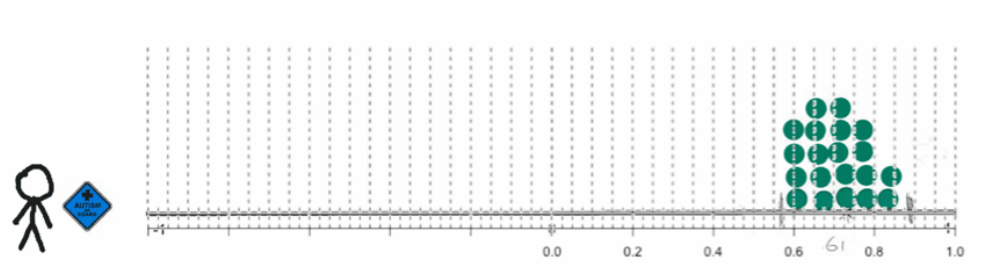


Fig. A.2.1: Distribution 1 for expert 1 with 80-bar grid (Step 1) applied.

The annotated R-code matching the digitalization steps for the first distribution of expert 1 is:

```
#general settings
f = 80
b <- seq(1,f,by=1)
pb <- vector(length=f) #expert specific values k=5
#step 1
#step 2
LL <- c(64,64,64,64,66)
UL <- c(75,75,72,72,70)
TP <- c(0.25,0.25,0.20,0.20,0.10)
N <- c(12,12,9,9,5) #N <- 1+(UL-LL)
#step 3: for every bar,
#allocate probability from each associated layer for (i in 1:f){
  for (l in 1:k){
    if (b[i] >= LL[l] & b[i] <= UL[l]){
    pb[i] = pb[i] + TP[l]/N[l]}
    }}
```

Histograms for digitalized trial roulette priors can be generated with the package LearnBayes (Albert, 2014) as follows:

```
midpt = seq(from=-1+2/f/2, to=1-2/f/2, by = 2/f)
p = seq(-1,1,length=2000)
plot(p,histprior(p,midpt,pb),type="l")
```

The result of this code is Figure A.2.2.

A.3 Derive Parametric Priors from Histogram Priors

Prior distributions and hyperparameters matching the digitalized judgment distributions (Appendix A.2) for the correlation can be found with the SHELF script. The function that should be used is:

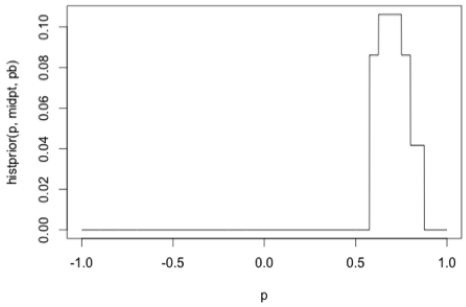


Fig. A.2.2: Correlation between cognitive potential and educational performance for youth with ASD according to expert 1.


```
elicit.group.values(N.experts = 4, method = "rp")
```

The function calls a window in which probability can be assigned to bars for the multiple experts. The probabilities of the each expert's vector pb were directly implemented. By making small adjustments to the SHELF script (nbins = 40 instead of 10), we were able to use 40 bars equally distributed over the interval (0,1), which was appropriate in our situation, because all experts indicated $p_b = 0$ for each bar in the interval (-1,0). The resulting priors for the four experts and two populations were:

$$p_{1\text{ASD}}(\rho) = \ln N(\rho| - 0.354, 0.014) I_{\rho\in[0,1]} \tag{A.1}$$
$$p_{2\text{ASD}}(\rho) = \text{Beta}(\rho|8.521, 7.167) \tag{A.2}$$
$$p_{3\text{ASD}}(\rho) = \text{Beta}(\rho|14.973, 7.181) \tag{A.3}$$
$$p_{4\text{ASD}}(\rho) = \ln N(\rho|0.708, 0.007) I_{\rho\in[0,1]} \tag{A.4}$$
$$p_{1\text{noASD}}(\rho) = N(\rho|0.462, 0.015) I_{\rho\in[0,1]} \tag{A.5}$$
$$p_{2\text{noASD}}(\rho) = N(\rho|0.252, 0.002) I_{\rho\in[0,1]} \tag{A.6}$$
$$p_{3\text{noASD}}(\rho) = \text{Beta}(\rho|21.727, 34.259) \tag{A.7}$$
$$p_{4\text{noASD}}(\rho) = \Gamma(\rho|32.224, 66.332) I_{\rho\in[0,1]}, \tag{A.8}$$

where ASD refers to the population of youth with ASD, noASD refers to the population of youth with diagnoses other than ASD, N denotes a normal distribution with a mean and variance, Beta denotes a beta distribution with hyperparameters alpha and beta, and Γ denotes a gamma distribution with hyperparameters shape and rate.

A.4 Code to Construct a Pool of Histogram Priors

When the probabilities 'pb', as described in Appendix A.2 are separately stored in vectors, the vectors can be simply added to construct the pooled histogram. For example, when the vectors are called pb1ASD, pb2ASD, pb3ASD, pb4ASD, pb1noASD,

pb2noASD, pb3noASD, and pb4noASD, the pooled priors can be made with the following code:

```
pbASD <- pb1ASD + pb2ASD + pb3ASD + pb4ASD
pbnoASD <- pb1noASD + pb2noASD + pb3noASD + pbno4ASD
```

A histogram can then again be constructed with the package LearnBayes (Albert, 2014) as follows:

```
f = 80
midpt = seq(from=-1+2/f/2, to=1-2/f/2, by = 2/f)
p = seq(-1,1,length=2000)
plot(p,histprior(p,midpt,pbASD/sum(pbASD)),type="l")
```


In the last line of code where the plot is constructed, pbASD is divided by its sum (i.e., 400), to make the total integrate to 1 again.

A.5 Code to Construct a Pool of Parametric Priors

A figure of a parametric prior can be easily constructed with R-code specifying each density, its weight (e.g., 1 divided by the number of experts for equal priors), and subsequently adding the densities. With the parametric priors described in Appendix A.3, the code to create a figure of the pooled distributions would be as follows:

```
curve(1/4*dlnorm(x,-0.3543455,0.1163011) +
1/4*dbeta(x,8.520659,7.167028) +
1/4*dbeta(x,14.973041,7.181388) +
1/4*dnorm(x,0.70803613,0.08189341), ylab="")
curve(1/4*dnorm(x,0.4620081,0.12047800) +
1/4*dnorm(x,0.2520595,0.04674401) +
1/4*dbeta(x,21.72679,34.25861) +
1/4*dgamma(x,32.76447,scale=1/65.41338), ylab="")
```

The pool of parametric distributions can be constructed by sampling from the separate parametric distributions and combining the resulting data. The sampling from truncated distributions can be done with the R-package Runuran (Leydold and Hörmann, 2015). The resulting pooled distribution, however, cannot be used as a prior directly. How a pooled parametric prior should be specified to be updated at once is software dependent. Some options are to write a sampler (Gill, 2014), write a module to add to existing software (Wabersich and Vandekerckhove, 2013), specify the (log)likelihood of the pooled parametric prior in Stan (Stan Development Team, 2014), or use the zeroes trick (Ntzoufras, 2009) in other software like OpenBUGS or Just Another Gibbs Sampler (JAGS; Plummer 2013).

A.6 Bayesian Updating

A.6.1 The likelihood

The likelihood function of the model is given by:

$$p(x_1, y_1, \ldots, x_s, y_s|\mu_x, \mu_y, \sigma_x, \sigma_y, \rho) = \prod_{i=1}^{s} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Big[-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x_i-\mu_x}{\sigma_x}\Big)^2 + \Big(\frac{y_i-\mu_y}{\sigma_y}\Big)^2 - 2\rho\Big(\frac{x_i-\mu_x}{\sigma_x}\Big)\Big(\frac{y_i-\mu_y}{\sigma_y}\Big)\Big]\Big], \tag{A.9}$$

where s denotes the number of subjects. In this bivariate normal distribution, the variance-covariance matrix is decomposed by means of the separation strategy (Barnard et al., 2000). This decomposition allows us to put a prior on ρ. Another advantage of this decomposition is that when $\rho \in [-1, 1]$, the variance-covariance matrix is always invertible.

A.6.2 Prior distributions

In the analysis for the population with ASD, the parametric expert priors were the following:

$$p_1(\rho) = \ln N(\rho| -0.354, 0.014) I_{\rho\in[0,1]} \tag{A.10}$$
$$p_2(\rho) = \text{Beta}(\rho|8.521, 7.167) \tag{A.11}$$
$$p_3(\rho) = \text{Beta}(\rho|14.973, 7.181) \tag{A.12}$$
$$p_4(\rho) = \ln N(\rho|0.708, 0.007) I_{\rho\in[0,1]} \tag{A.13}$$

In the analysis for the population with diagnoses other than ASD, the parametric expert priors were:

$$p_1(\rho) = N(\rho|0.462, 0.015) I_{\rho\in[0,1]} \tag{A.14}$$
$$p_2(\rho) = N(\rho|0.252, 0.002) I_{\rho\in[0,1]} \tag{A.15}$$
$$p_3(\rho) = \text{Beta}(\rho|21.727, 34.259) \tag{A.16}$$
$$p_4(\rho) = \Gamma(\rho|32.224, 66.332) I_{\rho\in[0,1]} \tag{A.17}$$

In the equations, N denotes a normal distribution with a mean and variance, Beta denotes a beta distribution with hyperparameters alpha and beta, and Γ denotes a gamma distribution with hyperparameters shape and rate.

In addition to the prior distributions for the correlation, a joint prior for the nuisance parameters $(\mu_x, \mu_y, \sigma_x, \sigma_y)$ was specified:

$$\begin{aligned} p(\mu_x, \mu_y, \sigma_x, \sigma_y) = \\ N(\mu_x|75.0, 400, 0) I(\mu_x \in [45, 145])) \times \\ N(\mu_y|0.75, 0.50) I(\mu_y \in [0.0, 1.5])) \times \\ \Gamma(\sigma_x|2.0, \frac{1}{7.5}) \times \Gamma(\sigma_y|2.0, 5.5) \end{aligned} \tag{A.18}$$

Justifications of the hyperparameters can be found in the main text, Section 4.2.3.

A.6.3 Posterior distribution

The posterior distribution is proportional to the prior times the likelihood of the data. The equation demonstrates this with a pooled expert prior.

$$p(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho | x_1, y_1, ..., x_s, y_s) \propto$$
$$p(\mu_x, \mu_y, \sigma_x, \sigma_y) \times \sum_{e=1}^{4} \left(\frac{1}{4} p_e(\rho) \right) \times p(x_1, y_1, ..., x_s, y_s | \mu_x, \mu_y, \sigma_x, \sigma_y, \rho). \quad \text{(A.19)}$$

The summation symbol can be moved, since it only sums over elements with subscript e, giving:

$$p(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho | x_1, y_1, ..., x_s, y_s) \propto$$
$$\sum_{e=1}^{4} \left(p(\mu_x, \mu_y, \sigma_x, \sigma_y) \times \frac{1}{4} p_e(\rho) \times p(x_1, y_1, ..., x_s, y_s | \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) \right). \quad \text{(A.20)}$$

These equations demonstrate that updating a pooled prior with the likelihood of the data is equal to pooling the posteriors of four analysis in which only one element of the pooled distribution was updated with the likelihood of the data.
The posterior was thus obtained for each expert separately. From each posterior, parameters could be sampled by means of an iterative procedure, which could then be combined according to the linear pooling principle, resulting in a posterior distribution.

A.6.4 Analysis

All Bayesian analyses were conducted in JAGS (Plummer, 2013) via the package rjags (Plummer, 2015) in R (R Core Team, 2015). In JAGS, a Gibbs sampler is used to approximate the posterior. The number of chains in each analysis was 3. Each of the chains consisted of 5,000 burn-in iterations, and 50,000 post burn-in iterations. JAGS' automatic random number generators, and seed values were adopted. As starting values for the chains, the maximum likelihood estimates were provided. Annotated R-code is provided as online supplementary material (Part IV).

Convergence of the analyses was assessed by inspecting the trace plots, and evaluating the potential scale reduction (PSR; Gelman and Rubin 1992). For the population with ASD, the convergence plots looked satisfactory for all posterior distributions. In addition, the PSR for the correlation parameters was calculated for every 100 iterations. For each expert, the PSR was lower than 1.05 in more than 97.8% of the evaluations. For more than 56.0% of the evaluations, the PSR was lower than 1.01. For the population with diagnoses other than ASD, the convergence plots looked satisfactory for all posterior distributions. In addition, the PSR for the correlation parameters was calculated for every 100 iterations. For each expert, the PSR was lower than 1.05 in more than 99.8% of the evaluations. For more than 65.4% of the evaluations, the PSR was lower than 1.01.

Part II

Testing Replication

5

Testing ANOVA Replications by Means of the Prior Predictive p-Value

Summary. In the current study, we explain how replication of an analysis of variance can be tested with the prior predictive p-value. That is, we test to what degree the new data deviates from data that can be predicted based on the original results, considering relevant features of the original study. The central role of claims by the original study is one of the unique features of the proposed method. These claims can, for example, concern specific values for the group means, the ordering of the group means, or effect sizes for between group differences.

We explain the calculation of the prior predictive p-value step by step, illustrate the method with examples, and elaborate on the topic of power. The replication test and its integrated power and sample size calculator are made available as interactive applications. As such, the current study supports researchers that want to adhere to the call for replication studies in the field of psychology.

5.1 Introduction

New studies conducted to replicate earlier original studies are often referred to as replication studies. After the latest "crisis in confidence" in the field of psychology, the call to conduct replication studies is stronger than ever (Anderson and Maxwell, 2016; Asendorpf et al., 2013; Cumming, 2014; Earp and Trafimow, 2015; Ledgerwood, 2014; Open Science Collaboration, 2012, 2015; Pashler and Wagenmakers, 2012; Schmidt, 2009; Verhagen and Wagenmakers, 2014). As a result, methodology on conducting replication studies is increasingly receiving attention (see, for example, Anderson and

This chapter is under review as Zondervan-Zwijnenburg, M.A.J., Van de Schoot, R. & Hoijtink, H., (under review). Testing ANOVA Replications by Means of the Prior Predictive p-Value. *Meta-Psychology*. doi: 10.31234/osf.io/6myqh

Author contributions: MZ and HH were involved in the initial research design. MZ drafted and revised the article in collaboration with HH. MZ developed the interactive application, conducted the simulation studies, and conducted the analyses. RS provided additional feedback, and evaluated the interactive application.

Maxwell, 2016; Asendorpf et al., 2013; Brandt et al., 2014; Schmidt, 2009). There is, however, no standard methodology to determine whether a replication is successful or not (Open Science Collaboration, 2015).

The results of an original study are replicated when a new study corroborates the original findings. A common and intuitive method to assess whether a result is replicated is 'vote-counting'. Vote-counting is assessing whether the new effect is statistically significant and in the same direction as the significant effect in the original study (Anderson and Maxwell, 2016; Simonsohn, 2015). Consider a situation with two studies: the first is the original study, and the second study is the new study. The original study results in a Cohen's d (Cohen, 1988) of .30 with a p-value of .01. A new study finds a Cohen's d of .30 with a p-value of .07. The new study would then be considered a non-replication of the original result, despite the fact that the effect sizes are the same. If the new study would find a Cohen's d of .10 with a p-value of .04, the new result would be considered a replication of the original result. Vote-counting has serious shortcomings. First of all, it is a dichotomous evaluation that does not take into account the magnitude of differences between effect-sizes (Asendorpf et al., 2013; Simonsohn, 2015). Secondly, each of the effect sizes being significant does not imply that both effect sizes are the same, nor does one significant effect and one non-significant effect imply that both effects are different (Gelman and Stern, 2006; Nieuwenhuis et al., 2011). Stated otherwise, vote-counting does not formally test whether a result is replicated (Anderson and Maxwell, 2016; Verhagen and Wagenmakers, 2014). Thirdly, underpowered replication studies are less likely to replicate significance, which can lead to misleading conclusions (Asendorpf et al., 2013; Cumming, 2008; Hedges and Olkin, 1980; Simonsohn, 2015).

In the current study, we address the following replication research question: "Does the new study fail to replicate relevant features of the original study?". Table 5.1 shows how our research question and proposed method relate to other replication research questions and associated methods. Our method addresses a question similar to that in Anderson and Maxwell (2016); Verhagen and Wagenmakers (2014); Harms (2018a); Ly et al. (2018) and Patil et al. (2016), but now enables researchers to evaluate the replication of relevant features of the original study other than effect sizes as well. The bottom panel of Table 5.1 shows other replication research questions that will not be pursued in this paper. The reader interested in these questions, should consult the given references.

As mentioned before, a unique characteristic of our method is that it tests the replication of relevant features of the original study, instead of effect sizes only. These relevant features concern the claims made by the original ANOVA study. For example, the original study often presents a certain ordering in the group means. The original findings, however, may reflect false positives, or may be sensitive to (intentional or unintentional) subjectivity of the original researchers. Therefore, we provide a test that confronts the new study, which is often conducted from a more critical and objective perspective, with the claims made by the original study. We thus test whether the new study rejects replication of important features found in the original study. In this perspective, the role of the original and new study are

Table 5.1: Replication Research Questions and Methods to Address Them

Replication Research Question	Method	Setting	Reference
Does the new study fail to replicate relevant features of the original study?	Prior predictive p-value	t-test, ANOVA	Current study
Does the new study fail to replicate the effect size of the original study?	Confidence interval for difference in effect sizes	t-test, correlation	Anderson and Maxwell (2016)
	Prediction interval	correlation	Patil et al. (2016)
Does the new study replicate the effect size of the original study?	Equivalence test	t-test	Anderson and Maxwell (2016)
	Bayes factor	t-test	Verhagen and Wagenmakers (2014)
	Bayes factor	ANOVA	Harms (2018a)
	Bayes factor	BF models[a]	Ly et al. (2018)
Is the effect present or absent in the replication attempt?	Bayes factor	t-test, correlation[b]	Marsman et al. (2017)
Is Cohen's d in the population of a detectable size?	Telescope test	t-test[c]	Simonsohn (2015)
What is Cohen's d in the population?	Confidence interval for average effect size	t-test	Anderson and Maxwell (2016)
What is the effect size (corrected for publication bias) in the population?	Hybrid meta-analysis	t-test	Van Aert and Van Assen (2017b)

[a] All models for which a Bayes factor can be computed.
[b] The reconceptualization by Ly et al. (2018) generalizes to most common experimental designs.
[c] The telescope test is explained in the t-test setting, but applicable to any model for which a power analysis can be conducted.

not symmetric: one is the hypothesis generator, and the other is the hypothesis confirmator. The claims or relevant features of original studies will be captured in the form of informative hypotheses (Hoijtink, 2012), which are specified using equality and inequality constraints among the means of the ANOVA model. We propose to evaluate the replication of these hypotheses with the prior predictive p-value (Box, 1980).

The prior predictive p-value was not introduced to test replication. It was originally presented as a method to test whether the current data is unexpected given the prior expectations concerning the parameter values and the statistical model. A disadvantage of the prior predictive check to test model fit is that it is leaves undetermined whether the prior expectations about the parameter values or the model assumptions are incorrect. Hence, as a model test the prior predictive check has been replaced by the posterior predictive check (Gelman et al., 1996), which does not make prior assumptions about expected parameter values, but instead uses the posterior results.

With respect to testing replication, however, the prior predictive check is a good method for three reasons. First, instead of prior expectations, we use the posterior distribution of the model parameters given the original data as the prior distribution. Consequently, we have a well-founded and clear-cut prior. Second, the prior predictive check uses a distribution of datasets that are expected given the prior (i.e., the original study). This prior predictive distribution takes variation in both the original study and the replication study into account. A study replicates if the new dataset is drawn from the same population as the original dataset. To include this variation, parameter values are sampled from the prior distribution (i.e., the posterior distribution of the original dataset), and given each sampled set of parameter values the predicted

datasets are simulated, that is: the predicted data is allowed to show sampling variance. Consequently, the prior predictive distribution takes into account that findings in a new dataset - resulting from a replication attempt - may deviate from the original findings because of random variation instead of meaningful differences. Third, the prior predictive check uses a 'relevant checking function'. The relevant checking function can concern any relevant feature of a study. As we will explain in the current paper, we propose to include an informative hypothesis based on the original study in this relevant checking function. As a result, we can check whether the new study fails to replicate relevant features of the original study, while taking variation around both studies into account.

The goal of this paper is to explain how the replication of relevant features of original ANOVA studies can be tested. In the first section, we provide a step by step introduction of the prior predictive p-value. To make our method to calculate the prior predictive p-value and sample sizes for new studies easily accessible, we provide an interactive application through the Open Science Framework at `osf.io/6h8x3` as a web tool and as an R-package (`ANOVAreplication`, Zondervan-Zwijnenburg 2018). In the second section, we use the interactive software to apply the prior predictive check to three examples from the Reproducibility Project Psychology (Open Science Collaboration, 2012). Finally, we dedicate the third section of this paper to the topic of power.

5.2 Prior Predictive *p*-Value

The evaluation of the replication of an ANOVA study by means of the prior predictive p-value (Box, 1980) consists of three steps that will be explained below.

5.2.1 Step 1: Prior Predictive Distribution of the Data

The ANOVA model is given by:

$$\begin{aligned} y_{ijd} &= \mu_{jd} + \epsilon_{ijd} \\ \epsilon_{ijd} &\sim \mathcal{N}(0, \sigma_d^2), \end{aligned} \tag{5.1}$$

where y_{ijd} is observation $i = 1, ..., n_{jd}$ in group $j = 1, ..., J$ for dataset $d \in \{o, r, \text{sim}\}$, where o denotes the original data, r denotes the new data, and sim denotes simulated data, the latter will be introduced towards the end of this section. Furthermore, μ_{jd} is the mean of group j in dataset d, ϵ_{ijd} is the error term, and σ_d^2 is the pooled variance over all J groups.

The original ANOVA results can be summarized in the posterior distribution of the parameters: $g(\boldsymbol{\mu}_o, \sigma_o^2 | \boldsymbol{y}_o)$, where $\boldsymbol{\mu}_o = [\mu_{1o}, ..., \mu_{Jo}]$ and $\boldsymbol{y}_o$ includes all observations y_{ijo}:

$$g(\boldsymbol{\mu}_o, \sigma_o^2 | \boldsymbol{y}_o) \propto f(\boldsymbol{y}_o | \boldsymbol{\mu}_o, \sigma_o^2) h(\boldsymbol{\mu}_o, \sigma_o^2), \tag{5.2}$$

where the density of the data

$$f(\boldsymbol{y}_o|\boldsymbol{\mu}_o, \sigma_o^2) = \prod_{j=1}^{J} \prod_{i=1}^{n_{jo}} \frac{1}{\sqrt{2\pi}\sigma_o} e^{\frac{-(y_{ijo}-\mu_{jo})^2}{2\sigma_o^2}} \tag{5.3}$$

and the standard prior distribution,

$$h(\boldsymbol{\mu}_o, \sigma_o^2) \propto \frac{1}{\sigma_o^2}. \tag{5.4}$$

The prior distribution is uninformative, that is, the posterior distribution is completely determined by the original data.

To test whether new data is in line with the original results, we need to obtain datasets that are to be expected given the original data. The prior distribution for future parameters $h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) = g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o)$. Using this prior we simulate data $\boldsymbol{y}_{\text{sim}}$ that are to be expected given the results of the original study:

$$f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{y}_o) = \int f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) d\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2 = f(\boldsymbol{y}_{\text{sim}}), \tag{5.5}$$

where $f(\boldsymbol{y}_{\text{sim}})$ is the prior predictive distribution of the data. Note that $f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2)$ is the counterpart of Equation 5.3 for dataset sim instead of o. Datasets $\boldsymbol{y}_{\text{sim}}^t$ for $t = 1, ..., T$, where T denotes the number of samples from the posterior, are obtained by sampling $\boldsymbol{\mu}_{\text{sim}}^t, \sigma_{\text{sim}}^{2,t}$ from $h(\boldsymbol{\mu}_{\text{sim}}, \sigma_{\text{sim}}^2) = g(\boldsymbol{\mu}_o, \sigma_o^2|\boldsymbol{y}_o)$ (see Appendix A.1 for a Gibbs sampler), and subsequently simulating $\boldsymbol{y}_{\text{sim}}^t$ from $f(\boldsymbol{y}_{\text{sim}}|\boldsymbol{\mu}_{\text{sim}}^t, \sigma_{\text{sim}}^{2,t})$ (cf. Equation 5.3). Thus, $f(\boldsymbol{y}_{\text{sim}})$ consists of datasets that we would expect given the results of the original study. Datasets $\boldsymbol{y}_{\text{sim}}^t$ have sample sizes $n_{1r}, ..., n_{Jr}$, because the predicted data needs to be compared to the new data $\boldsymbol{y}_r$ that has sample sizes $n_{1r}, ..., n_{Jr}$.

The steps in the following sections elaborate how new data $\boldsymbol{y}_r$ can be compared to data $f(\boldsymbol{y}_{\text{sim}})$ that are to be expected given the original results with a relevant checking function. This relevant checking function encompasses a hypothesis H_0 that represents relevant features of $\boldsymbol{y}_o$ as will be introduced in the next section.

5.2.2 Step 2: Hypotheses and Test Statistic

With the prior predictive p-value we want to test whether the new study fails to replicate relevant features of the original study. Hence, we are not interested in the classical null hypothesis claiming that "nothing is going on" (i.e., $\mu_{1d} = ... = \mu_{Jd}$). Instead, H_0 is used to catch the claims of the original study. We will provide three ways to create hypotheses representing relevant features of the original study.

First, it may be of interest whether the means as found in the original study are replicated in the new study. This implies that $\mu_{1r} = \bar{y}_{1o}, ..., \mu_{Jr} = \bar{y}_{Jo}$. For example, if $\bar{y}_{1o} = 1, \bar{y}_{2o} = 2, \bar{y}_{3o} = 3$ is observed, the corresponding hypothesis for the new study is H_0: $\mu_{1r} = 1, \mu_{2r} = 2, \mu_{3r} = 3$.

Second, in some situations it is of greater interest to see whether a new study corroborates more qualitative conclusions of the original study. A typical qualitative conclusion in the context of ANOVA studies is that the group means follow a certain

ordering. For example, a researcher may claim on the basis of the original results that the control group has a lower mean score than a group that received training A, which in turn has a lower mean score than the group that received training B: H_0: $\mu_{\text{control},r} < \mu_{\text{trainingA},r} < \mu_{\text{trainingB},r}$. Testing an ordering of means, implies for each pair j, j': $\mu_{jr} > \mu_{j'r}$, or $\mu_{jr} < \mu_{j'r}$, or $\mu_{jr}, \mu_{j'r}$. Another example is an original study showing that $\bar{y}_{1o} < \bar{y}_{3o}$ & $\bar{y}_{2o} < \bar{y}_{3o}$. Accordingly, the hypothesis for the new study is H_0: $\mu_{1r} < \mu_{3r}$ & $\mu_{2r} < \mu_{3r}$.

Third, it may be of interest to test if the effect sizes found in the original study can be replicated. Effect sizes can be quantified using Cohen's d. Using this implies that for one or more pairs j, j': Cohen's $d_{jj'r} = \frac{\mu_{jr} - \mu_{j'r}}{s_{jj'r}} \geq x$, where $s_{jj'r} = \frac{(n_{jr}-1)s^2_{jr} + (n_{j'r}-1)s^2_{j'r}}{n_{jr}+n_{j'r}-2}$. Furthermore, x denotes the minimum effect size. Given that we are testing more qualitative conclusions of the original study, we advise to place the original effect size in the qualitative categories as defined by Cohen (i.e., .00-.20 = negligible, 20-.50 = small, .50-.80 = medium, >.80 = large) and test the replication of an effect size of at least the size of the lower boundary of that category. For example, the original study may demonstrate that $\hat{d}_{12o} = .67$, and $\hat{d}_{23o} = .34$, where $\hat{d}_{jj'o}$ denotes an estimate of Cohen's d based on the observed data. Researchers may consider this result replicated if H_0: $d_{12r} \geq .5$ & $d_{23r} \geq .2$ is supported by the new data. This type of hypothesis is in line with Simonsohn (2015), who highlights the relevance of detectability of an effect with the example of levitation. If the original study documents 9 inch of levitation in an experimental group, researchers may be more interested in rejecting the qualitative claim that levitation is an existing and detectable phenomenon, that is testing, for example, $H_0 : d_{\text{control,experimental},r} > 0.2$, than in assessing whether a levitation of 7 inch as found in a new study is significantly different from the 9 inch in the original study, that is, testing H_0: $\mu_{\text{experimental},r} = 9.00$.

Above we have provided three options to construct hypotheses summarizing relevant features of the original study. These hypotheses concern mean values, order restrictions, and effect sizes. Which exact features should be covered in H_0 is to be decided based on the results and claims of the original study. The researcher conducting the replication test should substantiate the choices made in the formulation of H_0 with results from the original study. It is good practice to also pre-register H_0. In a later section of this paper we provide an example for each of the three types of hypotheses, but first we will explain how these hypotheses can be evaluated.

An F statistic that can quantify misfit for hypotheses like the H_0's introduced above is $\bar{F}$ (Silvapulle and Sen, 2005, p. 38-39):

$$\bar{F}_{\boldsymbol{y}_d} = \frac{\text{RSS}_{d,H_0} - \text{RSS}_{d,H_u}}{S^2_d}, \tag{5.6}$$

where RSS_{d,H_u} denotes the residual sum of squares in dataset $d \in \{r, \text{sim}\}$ for the unrestricted hypothesis H_u: $\mu_{1d}, ..., \mu_{Jd}$,

$$\text{RSS}_{d,H_u} = \sum_{ij} (y_{ijd} - \bar{y}_{jd})^2, \tag{5.7}$$

where $\bar{y}_{jd}$ denotes the mean for group j in dataset d. S_d^2 denotes the mean squared error,

$$S_d^2 = \frac{\text{RSS}_{d,H_u}}{N - J}, \tag{5.8}$$

where $N = \sum_{j=1}^{J} n_{jr}$, and

$$\text{RSS}_{d,H_0} = \sum_{ij} (y_{ijd} - \tilde{\mu}_{jd})^2, \tag{5.9}$$

where

$$\tilde{\boldsymbol{\mu}}_d = [\tilde{\mu}_{jd}, ..., \tilde{\mu}_{Jd}] = \underset{\tilde{\boldsymbol{\mu}}_d \in H_0}{\text{argmin}} \sum_{ij} (y_{ijd} - \mu_{jd})^2. \tag{5.10}$$

$\tilde{\boldsymbol{\mu}}_d$ thus contains the set of parameter estimates that minimize the residual sum of squares for $\boldsymbol{y}_d$ under the constraints imposed by H_0. $\bar{F}_{\boldsymbol{y}_d}$ is the scaled difference between the residual sum of squares under the constraints imposed by H_0 and the residual sum of squares for $\boldsymbol{y}_d$ under the unrestricted hypothesis H_u.

When we calculate $\bar{F}_{\boldsymbol{y}^t_{\text{sim}}}$ for each dataset $\boldsymbol{y}^t_{\text{sim}}$ obtained in Step 1 with respect to the hypothesis of interest from Step 2, a discrete representation of the prior predictive distribution of the test statistic $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is obtained. In the next section this distribution is used to compute the prior predictive p-value.

5.2.3 Step 3: p-value

The third and final step is to compute the prior predictive p-value.

$$p = P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r} | H_0) = \frac{1}{T} \sum_{t=1}^{T} I(\bar{F}_{\boldsymbol{y}^t_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r}), \tag{5.11}$$

where I is an indicator function that takes on the value 1 if the argument is true and 0 otherwise.

As illustrated in Figure 5.1, the prior predictive p-value indicates how exceptional the observed statistic for the new data, $\bar{F}_{\boldsymbol{y}_r}$, is compared to its prior predictive distribution $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$. The shaded area on the right side of $\bar{F}_{\boldsymbol{y}_r}$ is $P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} \geq \bar{F}_{\boldsymbol{y}_r} | H_0)$, that is, the prior predictive p-value. If the prior predictive p-value is significant, we reject replication of the relevant features of the original study by the new data.

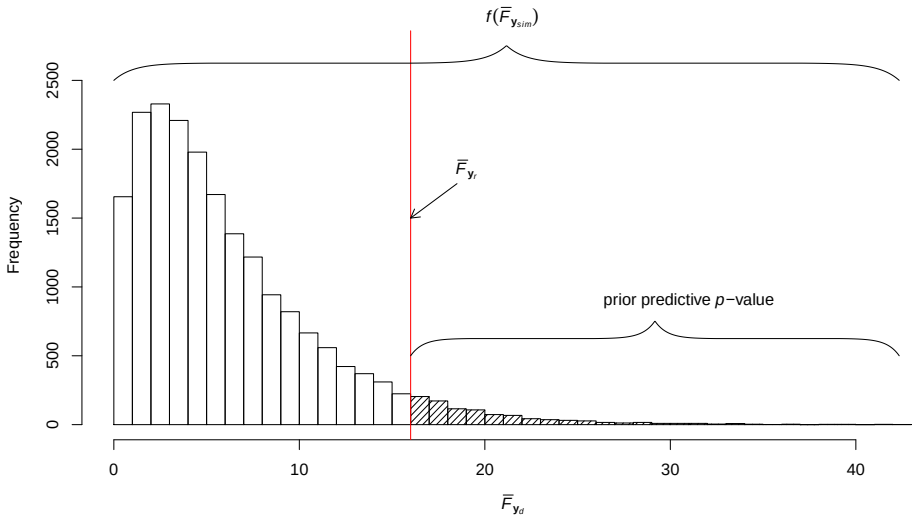


Fig. 5.1: An illustration of the prior predictive p-value.

Uniformity

To determine the significance of a p-value by comparing it to some preselected value α, the p-value needs to be uniformly distributed if replication is true. Only when the p-value is uniform, α is equal to the nominal Type I error. We will demonstrate that this is true for the prior predictive p-value and discuss two situations.

A p-value is uniform if:

$$f(p \leq \alpha | H_0) \leq \alpha \text{ for all } \alpha \in [0, 1], \tag{5.12}$$

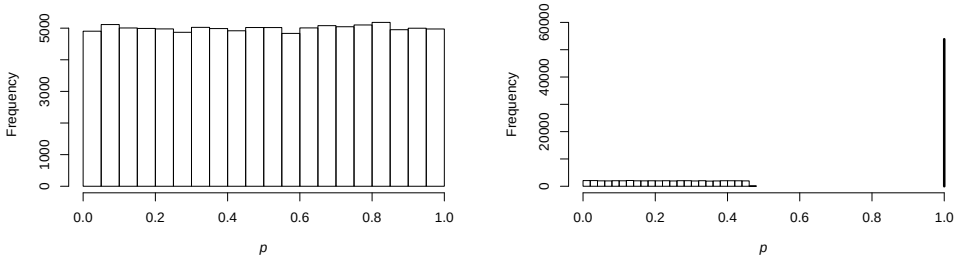
where p denotes a p-value from $f(p|H_0)$, that is, the null-distribution of the p-values. In Appendix A.2 it is proven that the prior predictive p-value is uniform if $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is continuous.

To illustrate our proof, consider the following situation:

- An original study with $\bar{y}_{1o} = 1, \bar{y}_{2o} = 2, \bar{y}_{3o} = 3, s_o^2 = 5$, and $n_{jo} = 50$.
- $H_0 : \mu_{1r} = 1, \mu_{2r} = 2, \mu_{3r} = 3$.
- $n_{jr} = 50$.

Following Step 1 and 2 of the prior predictive check we obtain $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$. Subsequently, we simulate $\boldsymbol{y}_r^t$ for $t = 1, ..., 100,000$ given $\mu_{1r} = 1, \mu_{2r} = 2, \mu_{3r} = 3, \sigma_r^2 = 5$, and calculate the prior predictive p-value for each $\boldsymbol{y}_r^t$, resulting in $f(p|H_0)$, which is plotted in Figure 5.2a. As can be seen in Figure 5.2a, $f(p|H_0)$ is nicely uniform.

When $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is discrete, however, the prior predictive p-value is not uniform for all $\alpha \in [0, 1]$. For example, when we use the same setup and procedure as in the previous paragraph, but now with $H_0 : \mu_{1r} < \mu_{2r} < \mu_{3r}$, we obtain Figure 5.2b, where the thick vertical line indicates a set of p-values with exactly the same value, namely 1.00. This set of equal p-values results from the fact that $H_0 : \mu_{1r} < \mu_{2r} < \mu_{3r}$ is true for a substantial number of datasets $\boldsymbol{y}_r^t$ causing the associated $\bar{F}_{\boldsymbol{y}_r^t}$ to be exactly equal to 0 and the associated prior predictive p-values to be exactly equal to 1. Generally, however, there exists an α_0 for which $f(p|H_0)$ is uniform (Meng, 1994), since all values in $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ other than 0 will occur in a continuous fashion. Thus, α is uniform for $\alpha \in [0, \alpha_0]$. For the discrete $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ considered here $\alpha_0 = .47$. For other situations, α_0 will equal $1 - P(f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}) = 0)$. A visualization of $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ can help to roughly estimate α_0. If the preselected $\alpha < \alpha_0$, α is equal to the nominal type I error.



(a) $f(p|H_0)$ for H_0: $\mu_{1r} = 1, \mu_{2r} = 2, \mu_{3r} = 3$. (b) $f(p|H_0)$ for H_0: $\mu_{1r} < \mu_{2r} < \mu_{3r}$.

Fig. 5.2: Illustrating the uniformity of the prior predictive p-value for two hypotheses.

Now that we have explained the three steps to compute the prior predictive p-value, we will demonstrate the application of the prior predictive p-value in the next section with three examples.

5.3 Examples of Testing Replication with the Prior Predictive p-Value

To illustrate the use of the prior predictive check to assess whether relevant ANOVA features are replicated, we selected replication studies that were part of the Reproducibility Project Psychology initiated by the Open Science Collaboration 2012; 2015. All calculations can be performed with the interactive application, which has been developed in Shiny (Chang et al., 2017) and the functions in our `ANOVAreplication` R-package (Zondervan-Zwijnenburg, 2018). The interactive application and R (R Core

Team, 2016) scripts to replicate the analyses for the running examples are all accessible through `osf.io/6h8x3`.

The first study is Janiszewski and Uy (2008), who study numerical judgements. More specifically, they study the impact of precision of an anchor, and motivation to adjust from the anchor on judgement bias. Janiszewski and Uy (2008), state: "Five studies show that adjustment away from a numerical anchor is smaller if the anchor is precise than if it is rounded" (p. 121). For the fourth experiment, this observation translates to H_0: ($\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r}$) & ($\mu_{\text{high motivation,round},r} > \mu_{\text{high motivation,precise},r}$). The resulting prior predictive p-value is 1.00. The data obtained by Chandler (2015) were perfectly in line with the H_0 describing the effect as observed by Janiszewski and Uy (2008). Therefore, we we conclude that the results of Janiszewski and Uy (2008) with respect to H_0: ($\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r}$) & ($\mu_{\text{high motivation,round},r} > \mu_{\text{high motivation,precise},r}$) are replicated by Chandler (2015).

Chandler (2015) replicated the study. Dr. Janiszewski provided dr. Chandler with the materials used in the original study, and both studies sampled college students. Overall, the new study is a close replication of the original study. Janiszewski and Uy (2008) did not draw conclusions about the exact mean values, but here we want to illustrate a hypothesis that tests the replication of exact values, that is, H_0: $\mu_{\text{low motivation,precise},r} = -0.76, \mu_{\text{low motivation,round},r} = -0.23, \mu_{\text{high motivation,precise},r} = -0.04, \mu_{\text{high motivation,round},r} = 0.98$. The resulting prior predictive p-value is $< .001$, indicating that the observed new data by Chandler (2015) obtains an extreme $\bar{F}$ score with respect to H_0 compared to the predicted data. Hence, for this H_0, we would reject replication. Again, note that the original study did not make these claims, hence, the rejection of its replication is not meaningful.

The group means, standard deviations, and sample sizes for the original study by Janiszewski and Uy (2008) and the replication attempt by Chandler (2015) are provided in Table 5.2.

Table 5.2: Z-scores of Participants' Mean Estimates from the Original Study: Janiszewski and Uy (2008), and the New Study: Chandler (2015)

	Low Motivation to Adjust				High Motivation to Adjust			
	Precise Anchor		Rounded Anchor		Precise Anchor		Rounded Anchor	
Study	*n*	*M (SD)*	*n*	*M (SD)*	*n*	*M (SD)*	*n*	*M (SD)*
Original	14	-0.76 (0.17)	15	-0.23 (0.48)	15	-0.04 (0.28)	15	0.98 (0.41)
New	30	-0.35 (0.23)	30	-0.18 (0.37)	30	0.20 (0.34)	30	0.35 (0.44)

Figures 5.3-5.6 show consecutively in the interactive application how: (1) the descriptive statistics of Janiszewski and Uy (2008) were entered to generate data with corresponding descriptives, (2) the posterior distribution to summarize the original data were obtained, (3) the new data of Chandler (2015) were uploaded, and (4) the results of the prior predictive check were acquired. In every figure, the input is

provided at the left hand side, while the output is provided at the right hand side. In Figure 5.3, the original data could have been uploaded instead of described if the data would have been available. Similarly, in Figure 5.5, the descriptive statistics of the new study, instead of the data, would have sufficed as well. The histogram in Figure 5.6 depicts the predictive distribution based on the data of Janiszewski and Uy (2008), while the red vertical line depicts $\bar{F}_{\boldsymbol{y}_r}$ for the data of Chandler (2015). Furthermore, the output in text provides a summary of $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$, the value of $\bar{F}_{\boldsymbol{y}_r}$, and the associated prior predictive p-value.

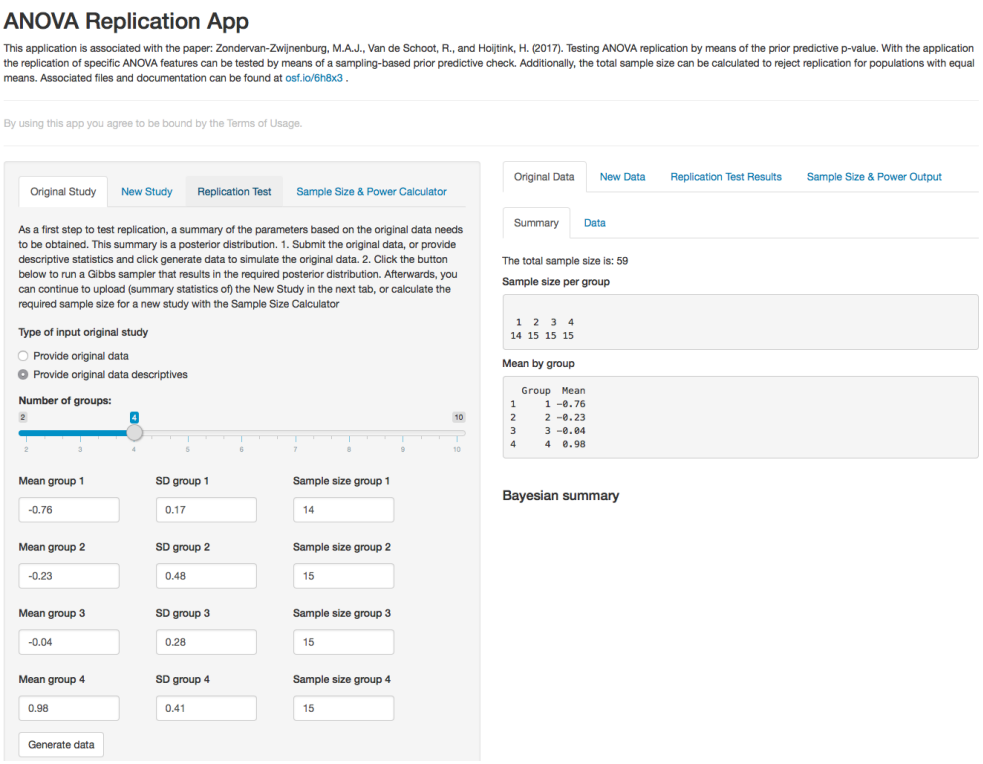


Fig. 5.3: Entering summary statistics of the original study in the interactive application.

The second study is Fischer et al. (2008), who studied the impact of self-regulation resources on confirmatory information processing. According to the theory, people who have low self-regulation resources (i.e., depleted participants) will prefer information that matches their initial standpoint. An ego-threat condition was added, because the literature proposes that ego-threat affects decision relevant information processing, although the direction of this effect is not clear. Fischer et al. (2008, p. 386) propose "[...] that participants with reduced self-regulation resources exhibit more pronounced

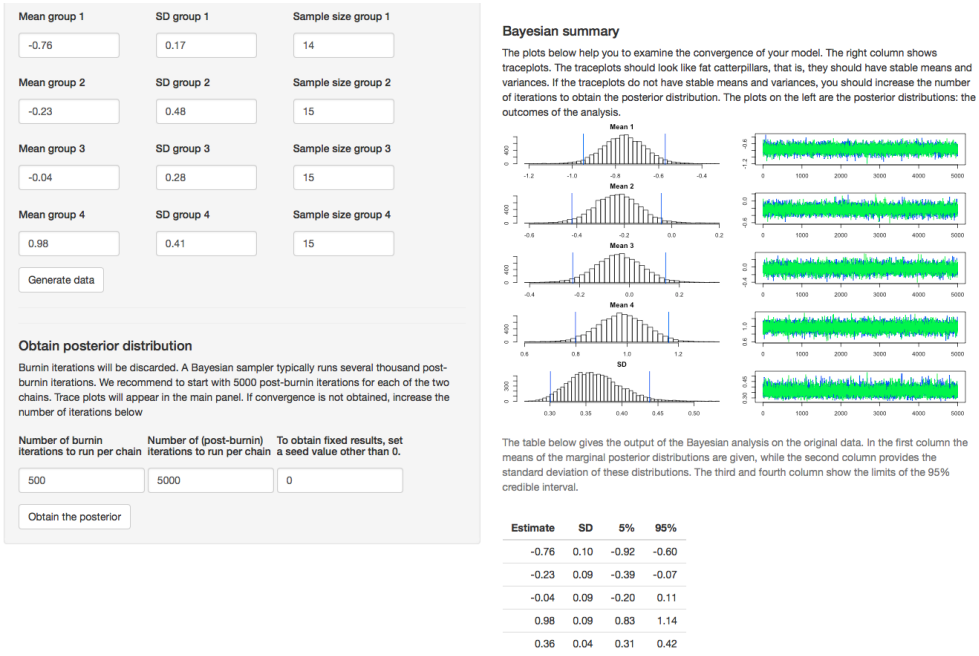


Fig. 5.4: Gibbs sampler in the interactive application.

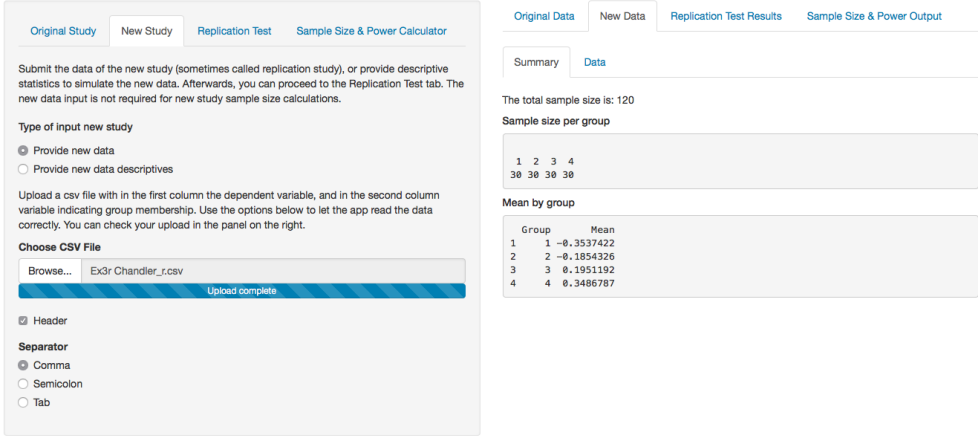


Fig. 5.5: Uploading replication data from a .csv file in the interactive application.

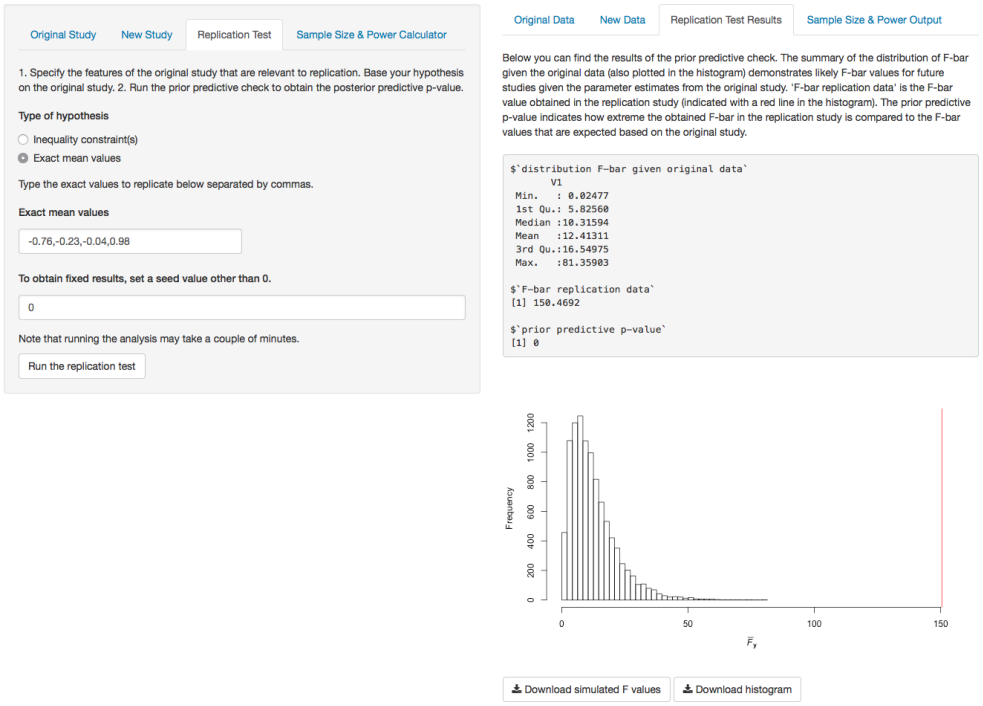


Fig. 5.6: Replication test input and output to test the replication of Janiszewski and Uy (2008).

confirmatory information processing than do nondepleted and ego threatened participants". The results for the dependent variable 'confirmatory information processing' are in line with this hypothesis. Consequently, with respect to a new study we would want to test H_0: $\mu_{\text{low self-regulation},r} > (\mu_{\text{high self-regulation},r}, \mu_{\text{ego-threatened},r})$. The group means, standard deviations, and sample sizes for the original study by Fischer et al. (2008) and the replication attempt by Galliani (2015) are provided in Table 5.3. The resulting prior predictive p-value is .003, indicating that we reject replication of H_0. The ordering in the new data by Galliani (2015) results in an extreme score $\bar{F}$ compared to the predicted data.

The third study is Monin et al. (2008), who studied the rejection of 'moral rebels'. The theory is that people who obey the status quo dislike rebels (as opposed to obedient others), because their own behavior is implicitly questioned by them. People who have been secured in their moral and adaptive adequacy by means of a self-affirmation task, however, should feel less need to reject rebels, and should feel able to recognize the value of the rebels' stand. With respect to the experiment that was subject to replication, Monin et al. (2008, p. 78) provide the following hypotheses: "Prediction 1a: Rejection by actors. Actors should like rebels less than they like obedient others.", and

Table 5.3: Descriptive Statistics for Confirmatory Information Processing from the Original Study: Fischer et al. (2008), and the New Study: Galliani (2015)

	Low self-regulation		High self-regulation		Ego-threatened	
Study	n	M (SD)	n	M (SD)	n	M (SD)
Original	28[a]	0.36 (1.08)	28[a]	-0.19 (0.53)	28[a]	-0.18 (0.81)
New	48	-0.07 (0.45)	47	-0.05 (0.47)	45	0.13 (0.64)

[a]Only the total sample size of 85 was provided in Fischer et al. (2008).

"Prediction 3a: Self-affirmation opens the heart. Self-affirmed actors should not feel a need to reject rebels as much as individuals less secure in their sense of self-worth, even if they still believe that rebels would dislike them." Monin et al. (2008) indeed observed for the dependent variable attraction: $\bar{y}_{\text{rebel},o} < (\bar{y}_{\text{rebel-affirmed},o}, \bar{y}_{\text{obedient},o})$. Furthermore, Monin et al. (2008) report that Cohen's $d_{\text{obedient,rebel},o}$ was .93 in this specific experiment, and that it was on average .86 over the four experiments that were part of the study. By Cohen's effect size categories, we could say that we replicate this study if we find that $d_{\text{obedient,rebel},r}$ is large. Since $\bar{y}_{\text{rebel},o} < \bar{y}_{\text{rebel-affirmed},o}$, $d_{\text{rebel-affirmed,rebel},r}$ should at least be positive. Consequently, with respect to a new study, we would want to test H_0: $d_{\text{obedient,rebel},r} \geq .80$, $d_{\text{rebel-affirmed,rebel},r} \geq 0$. The resulting prior predictive p-value is .154. Thus, we cannot reject replication of H_0. The group means, standard deviations, and sample sizes for the original study by Monin et al. (2008) and the replication attempt by Frank and Holubar (2015) are provided in Table 5.4.

The pressing question that we did not answer so far is: Was there enough power to reject replication in the first place? In the next part of the paper, we discuss the topic of power with a simulation study, an explanation of the importance of power, and a description of the power and sample size calculator. Furthermore, we calculate power for the Reproducibility Project examples introduced above.

Table 5.4: Descriptive Statistics for Attraction from the Original Study: Monin et al. (2008), and the New Study: Frank and Holubar (2015)

	Obedient		Rebel self-affirmed		Rebel	
Study	n	M (SD)	n	M (SD)	n	M (SD)
Original	19	1.88 (1.38)	19	2.54 (1.95)	29	0.02 (2.38)
New	20	0.98 (1.20)	27	0.02 (1.88)	28	0.27 (1.72)

5.4 Power

Power is the probability to reject the null hypothesis (of replication) with a preselected α when the null hypothesis is not true. Researchers typically pursue a power of .80.

Let us denote power by γ.

$$\begin{aligned}\gamma &= P(p < \alpha | H_a), \\ &= P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}} | H_a),\end{aligned} \tag{5.13}$$

where H_a is the hypothesized alternative population for which replication of H_0 is to be rejected. Note that the populations that can qualify to reject replication are endless. The population used is determined by the theoretical context in which the replication test takes place. The population with $\mu_{1a} = ... = \mu_{Ja}$ is a special population that is generally considered to display a non-effect in ANOVA studies. Hence, $\mu_{1a} = ... = \mu_{Ja}$ seems a natural choice for the alternative population.

In post-hoc power analyses, statistical power to reject the null hypothesis is computed considering the observed effect - in the context of replication this is the effect in the new study - as the population effect for which the null should have been rejected. Post-hoc power analyses provide a biased estimate of true power (Yuan and Maxwell, 2005). The power analysis introduced here is not a post-hoc power analysis, because (1) the datasets generated under the null hypothesis of replication are predicted from the original study, and (2) the datasets generated under the alternative hypothesis are independent of the new study, instead, H_a is the effect that one generally would like to reject replication for in ANOVA studies (i.e., $\mu_{1a} = ... = \mu_{Ja}$). Consequently, when we calculate the required sample size for $\gamma = .80$, this sample size is also not post-hoc, because the observed new data is not involved in its computation.

5.4.1 Simulation Study

To illustrate the power of the prior predictive p-value, we conducted a simulation study in which we varied the effect size in the original study f_o, the sample size for the original study n_{jo}, the sample size for the replication study n_{jr}, and the hypothesis of interest H_0. Table 5.5 summarizes the setup of the simulation. The population values for the alternative population were $\mu_{1a} = \mu_{2a} = \mu_{3a} = \bar{y}_o = 0.00$, where $\bar{y}_o$ denotes the grand mean in the original data, and $\sigma_a^2 = s_o^2 = 1.00$. μ_{ja} and σ_a^2 could take on any value, but we chose $\bar{y}_o$ and s_o^2 respectively, since they represent 'realistic' estimates. For each cell in the simulation study, 20,000 samples were drawn from H_a and power was calculated according to Equation 5.13.

The simulation estimates of γ to reject replication with $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$ are provided in Table 5.6. As expected, power increases with increasing effect sizes, and increasing sample sizes of the original study. If f_o, and n_{jo} are substantial, power increases as a function of the specificity of n_{jr} and H_0 as well. For small effects and low n_{jo}, larger sample sizes for the new study and more specific H_0 only emphasize the noise in the the original study more, and do not lead to an increase in power. The results show that when the effect size in the original study is small (i.e., $f = .10$), testing replication is a futile exercise: power was $< .15$ in all evaluated cells. Thus, small effects in the original study will often lead to results that seem to replicate, even when $\boldsymbol{y}_r$ is a sample from a population with $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$. When the sample

Table 5.5: Power Simulation Setup

H_0		H_a
$n_{jo} \in \{20, 50, 100\}$		$n_{jr} \in \{20, 50, 100\}$
$s_o^2 = 1.00$		$\sigma_a^2 = s_o^2 = 1.00$
f_o [a]		
.10	$\bar{y}_{1o} = -0.12, \bar{y}_{2o} = 0.00, \bar{y}_{3o} = 0.12$	$\mu_{1a} = \mu_{2a} = \mu_{3a} = \bar{y}_o = 0.00$
.25	$\bar{y}_{1o} = -0.31, \bar{y}_{2o} = 0.00, \bar{y}_{3o} = 0.31$	
.40	$\bar{y}_{1o} = -0.49, \bar{y}_{2o} = 0.00, \bar{y}_{3o} = 0.49$	
H_{01}: $\mu_{1r} < (\mu_{2r}, \mu_{3r})$		
H_{02}: $\mu_{1r} < \mu_{2r} < \mu_{3r}$		
H_{03}: $d_{12} \geq \frac{1}{2}d, d_{23} \geq 0$ [b]		
H_{04}: $\mu_{1r} = \bar{y}_{1o}, \mu_{2r} = \bar{y}_{2o}, \mu_{3r} = \bar{y}_{3o}$		

[a] f as introduced by Cohen (1988, p. 274-275).

[b] The standardized range of population means Cohen's $d = 2f\sqrt{\frac{3J-1}{J+1}}$ (Cohen, 1988, p. 279).

size per group in the original study is 20, the same applies: none of the combinations of effect sizes f_o, sample sizes n_{jr}, and hypotheses resulted in power $\geq .80$. When n_{jo} is 50, power can be sufficient if $n_{jr} \geq 50$ and the effect size is large. These results demonstrate that imprecise estimates (i.e., large standard errors) in the original study substantially lower the probability to reject replication for $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$. The power of the prior predictive p-value, however, is not surprisingly low: for a classical ANOVA study with three groups with a sample size of 20 each, power is <.10, <.40, and <.80 for small, medium, and large effect sizes respectively; a result that was already pointed out in Cohen (1988, p. 313). The fact that sample sizes of 20 per group result in insufficient power to reject replication for samples from a population with equal means, emphasizes the importance of substantial sample sizes. Note that power levels off for H_{01} and H_{02} at .667, and .833 respectively. Under $\mu_{1a} = \mu_{2a} = \mu_{3a}$, H_{01}:$\mu_{1r} < (\mu_{2r}, \mu_{3r})$ is true in $\frac{1}{3}$ of the situations by chance. Consequently, power cannot exceed $1 - \frac{1}{3} = .667$. For H_{02}:$\mu_{1r} < \mu_{2r} < \mu_{3r}$, $\frac{1}{6}$ of the combinations under H_a is in line with replication by chance. Hence, power cannot exceed $1 - \frac{1}{6} = .833$. However, γ will increase with H_a population parameters that deviate more from the original study parameters than $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$, and $\sigma_a^2 = s_o^2$. The simulation study with $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$ presents lower boundaries of γ if the new effect comes from a population where, for example, the ordering of means is switched.

5.4.2 The Importance of Power in Replication Studies

Only reporting the prior predictive p-value is not enough. Underpowered original studies may result in non-significant prior predictive p-values leading to the incorrect conclusion that not rejecting replication implies replication. The distorting impact of underpowered studies on psychological science in general and inferences about replication in particular is an omnipresent problem that has been emphasized already

Table 5.6: Power

f_o	n_{jo}	H_{01}			H_{02}			H_{03}			H_{04}		
		n_{jr}=20	50	100	20	50	100	20	50	100	20	50	100
.1	20	0.03	0.01	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
.1	50	0.08	0.06	0.03	0.07	0.04	0.02	0.07	0.05	0.02	0.03	0.01	0.00
.1	100	0.11	0.11	0.11	0.09	0.09	0.08	0.10	0.10	0.09	0.05	0.04	0.03
.25	20	0.13	0.09	0.05	0.09	0.05	0.01	0.12	0.10	0.06	0.05	0.03	0.01
.25	50	0.26	0.32	0.36	0.21	0.27	0.26	0.25	0.34	0.41	0.16	0.27	0.38
.25	100	0.31	0.45	0.61	0.27	0.43	0.55	0.31	0.48	0.64	0.23	0.47	0.73
.4	20	0.33	0.38	0.40	0.25	0.25	0.22	0.33	0.45	0.51	0.27	0.43	0.56
.4	50	0.50	*0.67*	*0.68*	0.46	0.69	***0.83***	0.51	0.74	**0.90**	0.51	0.86	**0.99**
.4	100	0.56	*0.68*	*0.68*	0.55	***0.84***	***0.84***	0.58	**0.84**	**0.97**	0.62	**0.95**	**1.00**

Text in cells with $\gamma \geq .80$ is boldface.

Text in cells with a maximum γ is italic.

by Maxwell (2004) and many others. As we will demonstrate in the current section, we can overcome the problem of underpowered studies for the prior predictive check by not only reporting the prior predictive p-value, but also the power as defined in the previous section. On the other hand, studies can also be overpowered. We will discuss the issue of overpowered studies at the end of the current section.

An underpowered original study may, for example, have the following features: $\bar{y}_{1o} = -0.35, \bar{y}_{2o} = 0.00, \bar{y}_{3o} = 0.35, s_o^2 = 1, n_{jo} = 20$ for $j = 1, 2, 3$, reflecting a medium effect size ($f = .29$ Cohen, 1988, p. 274-275) in combination with a small sample. If in the new study $\bar{y}_{1r} = 0.10, \bar{y}_{2r} = 0.00, \bar{y}_{3r} = -0.10, s_r^2 = 1, n_{jr} = 100$, that is, the ordering in the new study is reversed compared to the ordering observed in the original study, and we test H_0: $\mu_{1r} < \mu_{2r} < \mu_{3r}$, we obtain a prior predictive p-value of .116 and cannot reject replication with $\alpha = .05$! As demonstrated by the simulation study (see for example the fourth row of Table 5.6) and the example above, the power to reject replication can be low when the sample size in the original study is small. This is as true for the prior predictive p-value as it is for other approaches. As highlighted by Patil et al. (2016): Replication can only be rejected based on the claims that the original study makes, and when these claims are vague, rejecting them is hard to impossible.

To address the issue of underpowered original studies leading to non-significant prior predictive p-values, our software also calculates power to reject replication when all group means are equal instead of in line with H_0. Note that H_0 here is an informative hypothesis, and not the traditional null hypothesis that "nothing is going on". Then, if a non-significant prior predictive p-value is obtained, replication can only be deemed successful if the corresponding power to reject replication in case of equal means was sufficient. In the example above resulting in a prior predictive p-value of .116, the

corresponding power was .03, warning us against interpreting the non-significant prior predictive p-value as an indication of replication. Moreover, using the sample size calculator, we learn that power for this example does not increase with larger sample sizes for the new study. Therefore, it has to be concluded that the precision of the original study was insufficient to test replication using new data.

If an original study has high statistical power (i.e., small standard errors) and the conclusions of the study focus on its specific mean values, then a study can be overpowered. That is, small and possibly irrelevant deviations between the original and new study means may cause replication to be rejected. For example, an original study may have the following features: $\bar{y}_{1o} = -0.35, \bar{y}_{2o} = 0.00, \bar{y}_{3o} = 0.35, s_o^2 = 1$. Then, for a large sample size replication is rejected if we test H_0: $\mu_{1r} = -0.35, \mu_{2r} = 0.00, \mu_{3r} = 0.35$, even for a new study with $\bar{y}_{1r} = 0.30, \bar{y}_{2r} = 0.00, \bar{y}_{3r} = -0.30, s_r^2 = 1$. This issue, however, will not occur if we test the replication of the constellation of means as found in the original study: H_0: $\mu_{1r} < \mu_{2r} < \mu_{3r}$. In the same vein, there is no issue if we test whether the effect sizes for pairs of means in the original study are recovered by the new study H_0: $d_{12r} \geq .20$ & $d_{13r} \geq .50$ & $d_{23r} \geq .20$. In both situations, replication will not be rejected due to irrelevant differences between means over the studies, but only if the new study substantially deviates with respect to H_0. Note that we do not advice to change the specification of H_0 based on power: study conclusions remain leading in determining H_0.

Thus, to tackle the impact of underpowered studies, our interactive application provides a power and sample size calculator. Overpowered studies will seldom be an issue, since the prior predictive p-value is only sensitive to possibly irrelevant differences between studies if the original study conclusions require the inclusion of specific mean values in H_0. Non-significant prior predictive p-values for well-powered replication tests indicate that H_0 replicates. Non-significant prior predictive p-values for underpowered replication tests may indicate that the original study is too imprecise to test replication of H_0, and/or that the sample size in the new study is too small to test replication. The sample size calculator can clarify if a larger new study would render a useful replication test, or if testing the replication of H_0 based on the original study is pointless, because the original study is not informative. By calculating the power of the prior predictive check, and in case of insufficient power, by calculating the required sample size for a new study, we can draw meaningful conclusions with respect to the replication and replicability of the original study.

5.4.3 Power and Sample Size Determination

As highlighted in the previous sections and in the literature (e.g., Brandt et al., 2014; Simonsohn, 2015), power is an important characteristic of a convincing replication study. It is thus important that researchers can calculate the power of the prior predictive check, and can determine the sample size for a new study such that the replication test has high statistical power. Therefore, the interactive application includes a power and sample size calculator.

To calculate power with the interactive application, the user is required to (1) upload $\boldsymbol{y}_o$ or to provide group means, standard deviations, and sample sizes for $\boldsymbol{y}_o$, (2) specify H_0, and (3) provide $\boldsymbol{n}_r$, where $\boldsymbol{n}_r$ is a vector of length J containing the sample size for each group j. Furthermore, the user can designate the significance level α, which is by default set to .05.

Given $\boldsymbol{y}_o$, H_0, $\boldsymbol{n}_r$, and α, the power γ is calculated as follows:

1. Following Step 1 and 2 of the prior predictive check with $\boldsymbol{y}_o$, H_0, $\boldsymbol{n}_r$, and α, $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$ is obtained, and $\bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}$ can be calculated.
2. $t = 1, ..., T$ datasets $\boldsymbol{y}_r^t|H_a$ are simulated with $\mu_{ja} = \bar{y}_o$, $\sigma_a^2 = s_o^2$, and $\boldsymbol{n}_r$. Following Step 2 of the prior predictive check given $\boldsymbol{y}_r^t|H_a$ and H_0, $\bar{F}_{\boldsymbol{y}_r^t}|H_a$ is calculated.
3. $\gamma = P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}|H_a) =$

$$\frac{1}{T}\sum_{t=1}^{T} I(\bar{F}_{\boldsymbol{y}_r^t} \geq \bar{F}_{\boldsymbol{y}_{\text{sim},1-\alpha}}),$$

To determine the required sample size to reject replication with sufficient power, the sample size calculator in the interactive application uses an iterative procedure. First, if not already done so, the user is required to (1) upload $\boldsymbol{y}_o$ or to provide group means, standard deviations, and sample sizes for $\boldsymbol{y}_o$, and (2) specify H_0. Furthermore, the user can provide a target power level $\tilde{\gamma}$; a margin for the target power γ_{margin}, because the calculated power may not be exactly equal to the target power; the significance level α; a starting value for the group sample size n_{jr_0}; a maximum number of iterations $Q_{\max}$; and a maximum total sample size for the new study $N_{r_{\max}}$. The default values are: $\tilde{\gamma} = .825$, $\gamma_{\text{margin}} = .025$, $\alpha = .05$, $n_{jr_0} = 20$, $Q_{\max} = 10$, and $N_{r_{\max}} = 600$.

Given $\boldsymbol{y}_o$, H_0, $\tilde{\gamma}$, γ_{margin}, α, n_{jr_0}, $Q_{\max}$, and $N_{r_{\max}}$, the required sample size for a new study to reject replication with sufficient power when H_a: $\mu_{1a} = ... = \mu_{Ja} = \bar{y}_o$, and $\sigma_a^2 = s_o^2$ is calculated as follows:

1. In every iteration q, γ_q is calculated given n_{jr_q}.
2. When $q > 1$, $n_{jr_{q+1}}$ is determined by regressing $\{\gamma_1, ..., \gamma_i\}$ on $\{n_{jr_1}, ..., n_{jr_q}\}$ with a linear or quadratic (only if $q = 3$) function. In case of a linear regression, the linear regression coefficient β_1 is the power increase per subject. Subsequently, $n_{jr_{q+1}} = (\gamma_q - \tilde{\gamma})/\beta_1 + n_{jr_q}$. In case of regression with a quadratic function, $n_{jr_{q+1}}$ is calculated by solving the polynomial: $\tilde{\gamma} = \beta_0 + \beta_1 n_{jr_{q+1}} + \beta_2^2 n_{jr_{q+1}}$.
3. Repeat step (1) and (2) until $\gamma_q \in [\tilde{\gamma} - \gamma_{\text{margin}}, \tilde{\gamma} + \gamma_{\text{margin}}]$ (i.e., power is sufficient), or $\gamma_{q-1} \approx \gamma_q$ (i.e., power does not increase anymore up to two decimal points), or $n_{jr_{q-1}} = n_{jr_q}$ (i.e, the sample size does not change anymore), or $q = Q_{\max}$, or $\Sigma_{j=1}^{J} n_{jr_q} = N_{\max}$.

In the next section we calculate the power for the examples, and demonstrate the sample size calculator.

5.4.4 Power in the Examples

For the replication of Janiszewski and Uy (2008) by Chandler (2015), we first tested H_0: ($\mu_{\text{low motivation,round},r} > \mu_{\text{low motivation,precise},r}$) & ($\mu_{\text{high motivation,round},r} >$

$\mu_{\text{high motivation,precise},r}$). The resulting prior predictive p-value was 1.00, and the associated power was .75. Although .75 may not qualify as sufficient power, it is the maximum proportion of power for this H_0, which is true in $\frac{1}{4}$ of the samples under H_a. The data obtained by Chandler (2015) were perfectly in line with the H_0 describing the effect as observed by Janiszewski and Uy (2008). Therefore, we we conclude that the results of Janiszewski and Uy (2008) are replicated by Chandler (2015). As an illustration, we also tested H_0: $\mu_{\text{low motivation,precise},r} = -0.76$, $\mu_{\text{low motivation,round},r} = -0.23$, $\mu_{\text{high motivation,precise},r} = -0.04$, $\mu_{\text{high motivation,round},r} = 0.98$. The resulting prior predictive p-value was $<.001$ with a power of 1.00. Thus, this illustrative H_0 did not replicate, but Janiszewski and Uy (2008) may have been overpowered to test the replication of exact values, and hence, the differences between the studies may not have been meaningful. Again, we emphasize that the original study conclusions should be leading in determining H_0.

For the replication of Fischer et al. (2008) by Galliani (2015), we tested H_0: $\mu_{\text{low self-regulation},r} > (\mu_{\text{high self-regulation},r}, \mu_{\text{ego-threatened},r})$. The resulting prior predictive p-value was .003. The associated power was .66, indicating that we reject replication, despite low power. Apparently, the results by Galliani (2015) deviate even more from Fischer et al. (2008) than 34% of the samples under H_a in which all means are equal.

For the replication of Monin et al. (2008) by Frank and Holubar (2015), we tested H_0: $d_{\text{obedient,rebel},r} \geq .80, d_{\text{rebel-affirmed,rebel},r} \geq 0$. The resulting prior predictive p-value was .154 with a power of .77. This result is also demonstrated by the interactive application in Figure 5.7. Thus, we cannot reject replication of H_0, but this may be caused by a lack of power. The sample size calculator (Figure 5.8) shows that the group sample size in a new study needs to be at least 28 per group to achieve sufficient power to reject replication. Since 28 per group seems a conceivable number, we consider the conclusion of Monin et al. (2008) a suitable candidate for replication testing, but it requires slightly larger sample sizes than currently obtained in Frank and Holubar (2015) to arrive at sharp conclusions.

5.5 Conclusion

The goal of the current paper was to introduce the prior predictive check as a manner to test replication of ANOVA features. Additionally, we developed an interactive application (see `osf.io/6h8x3`) that enables all researchers to make use of our contribution. With the prior predictive check researchers can find an answer to the question: "Does the new study fail to replicate relevant features of the original study?" Identifying a non-replication may make us wonder about the representativeness of the original study, the new study, and the comparability of both studies. Or, as stated by Simonsohn (2015, p. 9) "Statistical techniques help us identify situations in which something other than chance has occurred. Human judgment, ingenuity, and expertise are needed to know what has occurred instead."

In the current paper, we discussed the prior predictive p-value for the ANOVA setting. In this manner, we were able to elaborate on specific informative hypotheses, the

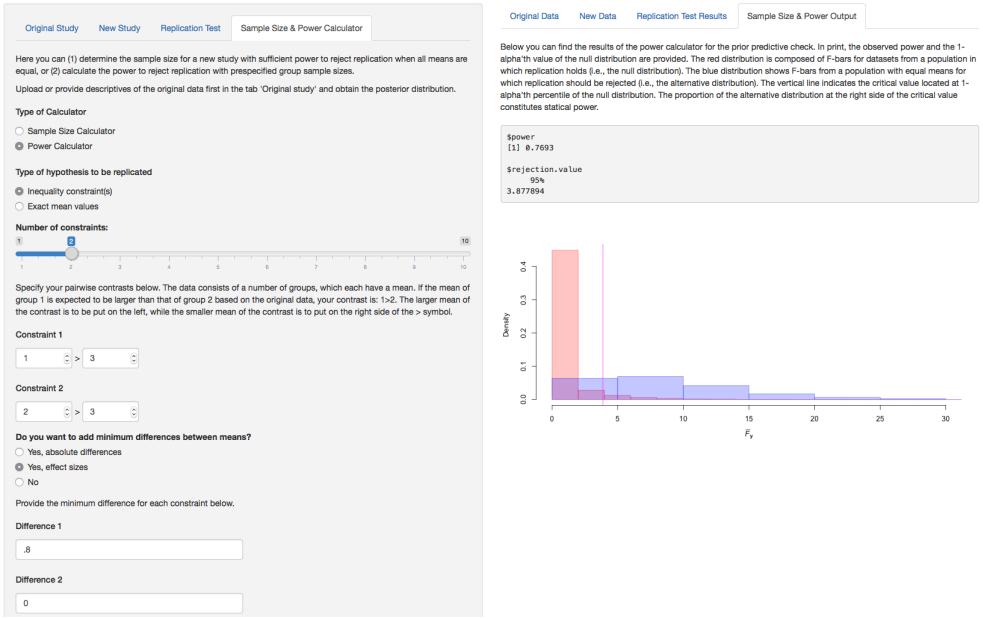


Fig. 5.7: Power calculator.

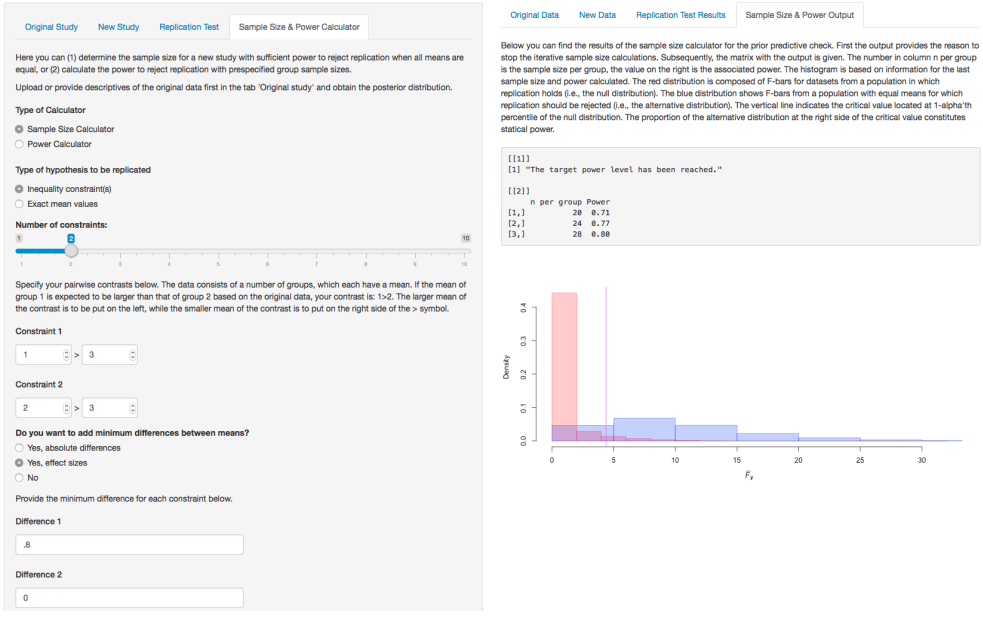


Fig. 5.8: Sample size calculator.

properties of the prior predictive p-value, and three examples from the Reproducibility Project Psychology. To test replication, the prior predictive p-value, however, is generalizable to statistical models other than the ANOVA as well. That is, for any model a predictive distribution can be obtained, informative hypotheses can be constructed, and a test-statistic evaluating the constraints can be calculated. The test as currently provided can already be used for the repeated measures ANOVA by means of contrast weights (see, for example, Furr and Rosenthal, 2003). With contrast weights a score for each participant can be calculated indicating to what degree the participant follows the expected pattern. Subsequently, the replication of relevant features of these contrast scores over groups can be tested.

By proposing the use of the prior predictive check in the context of replication, we provide researchers with an actual, and easy to use test for replication of ANOVA features. The availability of this test can further promote the trend to conduct more replication studies in the field of psychology.

A

Appendices

A.1 R Code to Sample from Posterior Distribution

Below we provide code which can be run in R (R Core Team, 2016) to sample from $f(\boldsymbol{\mu}_o, \sigma_o^2)$. We used a Gibbs sampler derived from Lynch (2007, p. 170-172).


```
# Gibbs sampler ####
Gibbs.ANOVA <- function(data,it=10000,burnin=500){
#R program for Gibbs sampling from conditionals

I=it+burnin
x=data$g; y=data$y
N=length(y); G=length(unique(data$g))

fit.lm <- lm(y~as.factor(x)-1)                    #lm, no intercept
x <- model.matrix(fit.lm)[,,drop = FALSE]

#establish parameter vectors and constant quantities
s1=matrix(1,I); b1=matrix(0,I,G)
s2=matrix(1,I); b2=matrix(0,I,G)
xtxi=solve(t(x)%*%x)
pars=coefficients(lm(y~x-1))
#Gibbs sampling begins
for(t in 2:I){ #Chain 1
  #simulate beta from its multivariate normal conditional
  b1[t,]=pars+t(rnorm(G,mean=0,sd=1))%*%chol((s1[t-1]^2)*xtxi)
  #choleski decomposition
  #simulate sigma from its inverse gamma distribution
  s1[t]=sqrt(1/rgamma(1,N/2,.5*t(y-x%*%(b1[t,]))%*%
                                (y-x%*%(b1[t,]))))
}
for(t in 2:I){ #Chain 2
  #simulate beta from its multivariate normal conditional
  b2[t,]=pars+t(rnorm(G,mean=0,sd=1))%*%chol((s2[t-1]^2)*xtxi)
```

```
  #choleski decomposition
  #simulate sigma from its inverse gamma distribution
  s2[t]=sqrt(1/rgamma(1,N/2,.5*t(y-x%*%(b2[t,]))%*%
                                (y-x%*%(b2[t,]))))
}

## store samples from both chains
par1=cbind(b1,s1)[-c(1:burnin),]
par2=cbind(b2,s2)[-c(1:burnin),]
results <<- rbind(par1,par2)
colnames(results) <- paste("Mean",1:(G+1))
colnames(results)[G+1] <- "SD"
```

A.2 Proof of Uniformity

The following three steps proof that Equation 5.12 holds for the prior predictive p-value when the distribution of the test statistic is continuous:

1. $P(p < \alpha|H_{0c}) = P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}|H_{0c})$, where $\bar{F}_{\boldsymbol{y}_r}$ is the test-statistic rendering p via $p = P(\bar{F}_{\boldsymbol{y}_{\text{sim}}} > \bar{F}_{\boldsymbol{y}_r}|H_{0c})$ and $\bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}$ is the 1-αth percentile of the distribution $f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}|H_{0c})$.
2. $P(\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}|H_{0c}) = \int_{\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_r}|H_{0c}) d\bar{F}_{\boldsymbol{y}_r}$, where $f(\bar{F}_{\boldsymbol{y}_r}|H_{0c})$ denotes the distribution of $\bar{F}_{\boldsymbol{y}_r}$ under H_{0c}.
3. For the situations considered in this paper it holds that $f(\bar{F}_{\boldsymbol{y}_r}|H_{0c}) = f(\bar{F}_{\boldsymbol{y}_{\text{sim}}})$, therefore $\int_{\bar{F}_{\boldsymbol{y}_r} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_r}|H_{0c}) d\bar{F}_{\boldsymbol{y}_r} = \int_{\bar{F}_{\boldsymbol{y}_{\text{sim}}} > \bar{F}_{\boldsymbol{y}_{\text{sim}},1-\alpha}} f(\bar{F}_{\boldsymbol{y}_{\text{sim}}}) d\bar{F}_{\boldsymbol{y}_{\text{sim}}} = \alpha$, which competes the proof.

6

How to Test Replication for Structural Equation Models

Summary. This paper introduces the prior predictive p-value as a manner to test replication in structural equation models. With the prior predictive p-value, the user tests whether the new study fails to replicate relevant original study findings as captured in the replication hypothesis. Using the replication of a piecewise latent growth model as a running example, the study explains the steps to obtain the prior predictive p-value and illustrates them with R-code. Finally, the study demonstrates how the replication of a more advanced structural equation model - a multilevel latent growth curve model - can be tested. All steps to compute the prior predictive p-value are also incorporated in the `Replication` R-package.

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 effect sizes. See for example, Anderson and Maxwell (2016); Harms (2018a); Ly et al. (2018); Patil et al. (2016). To test the failure to replicate relevant findings obtained with ANOVA models, one can use the method presented in Zondervan-Zwijnenburg et al. (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). 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 direction and significance of parameters is repeated in a second study (e.g., Carleton et al., 2010; Stokes et al., 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

This chapter will be submitted as Zondervan-Zwijnenburg, M.A.J. How to Test Replication for Structural Equation Models.

Author contributions: MZ is the sole author.

with samples from the same population due to sampling variance and measurement error (Patil et al., 2016; Stanley and 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 also explains 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 (Zondervan-Zwijnenburg, 2019). In the background, the `Replication` package uses `lavaan` (Rosseel, 2012) and `blavaan` (Merkle and 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 et al., 2004; Depaoli and Van de Schoot, 2017; Van de Schoot et al., 2013). As Supplementary Materials, R-code and data are provided that can be used to execute each of the steps in this paper. Note that for privacy considerations, the associated datasets in the Supplementary Materials at `https://osf.io/as7kz` are simulated data based on the covariance matrix of the original datasets. Consequently, the results can differ from those as reported 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.

For both examples, data and permission for their use were received from key researchers in the projects

6.1 The Prior Predictive *p*-value

Patil et al. (2016) and Zondervan-Zwijnenburg et al. (2019) introduced the idea that when we observe the results of an original study, it gives us expectations about results from future replication efforts. 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 deviation from 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 in terms of deviation from H_0 than the new observed data. This proportion is the prior predictive p-value. A small prior predictive p-value indicates that the results of the new observed data are 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.

6.1.1 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, 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 $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$.

$$g(\boldsymbol{\theta}_o|\boldsymbol{y}_o) \propto f(\boldsymbol{y}_o|\boldsymbol{\theta}_o)h(\boldsymbol{\theta}_o), \tag{6.1}$$

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:

$$\int f(\boldsymbol{y}_s|\boldsymbol{\theta}_s)h(\boldsymbol{\theta}_s)d\boldsymbol{\theta}_s = f(\boldsymbol{y}_s), \tag{6.2}$$

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 with $f(\boldsymbol{y}_s)$. The section "The Prior Predictive Distribution" elaborates on the procedure to obtain $f(\boldsymbol{y}_s)$.

6.1.2 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 and 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.

6.1.3 Step 3: The Prior Predictive *p*-value

Given H_0, we compute the test statistic D for the predicted and new data resulting in $f(D_{\boldsymbol{y}_s})$ and $D_{\boldsymbol{y}_r}$. 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 hypothesis H_u where $\boldsymbol{\theta}$ is estimated as usual to best fit the data at hand (Silvapulle and Sen, 2005, p. 59-63):

$$\begin{aligned} D &= \ln \frac{f_u}{f_0} \\ &= (\ln f_u - \ln f_0), \end{aligned} \tag{6.3}$$

where

$$f_u = \max_{\boldsymbol{\theta} \in H_u} f(\boldsymbol{\theta}|\boldsymbol{y}_d), \tag{6.4}$$

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

$$f_0 = \max_{\boldsymbol{\theta} \in H_0} f(\boldsymbol{\theta}|\boldsymbol{y}_d), \tag{6.5}$$

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 and 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 for data that we expect given the original results. Finally, we can compute the prior predictive p-value:

$$P(D_{\boldsymbol{y}_s} \geq D_{\boldsymbol{y}_r}|H_0). \tag{6.6}$$

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.

6.2 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).

```
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 ANOVA 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:

$$\boldsymbol{y} = \boldsymbol{\nu} + \boldsymbol{\Lambda\eta} + \boldsymbol{\epsilon}, \tag{6.7}$$

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:

$$\boldsymbol{\eta} = \boldsymbol{\alpha} + \boldsymbol{\zeta}, \tag{6.8}$$

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}, $\Theta_{\text{positive}}, \Theta_{\text{neutral}}$, and Θ_{negative}.

The `lavaan` model syntax is provided in Appendix B.1 and in the Supplementary R-code at `https://osf.io/as7kz`. 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 prior predictive p-value, 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.

6.3 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 $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$. The posterior $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$ is our prior for predicted data $h(\boldsymbol{\theta}_s)$. In this manner, we base the prior predictive distribution on the original results. The remainder of this section illustrates with R-code how a prior predictive distribution can be obtained. The data and code to reconstruct all output are provided in the Supplementary Materials at `https://osf.io/as7kz`.

The posterior distribution $g(\boldsymbol{\theta}_o|\boldsymbol{y}_o)$ 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. Below, we run a Bayesian analysis on the Achterberg et al. (2017) data with the R-package `blavaan` (Merkle and Rosseel, 2018) with the default prior distributions (see Appendix B.2) 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. Figure 6.1 shows the histogram that depicts the samples from the posterior for α_{i}. 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.

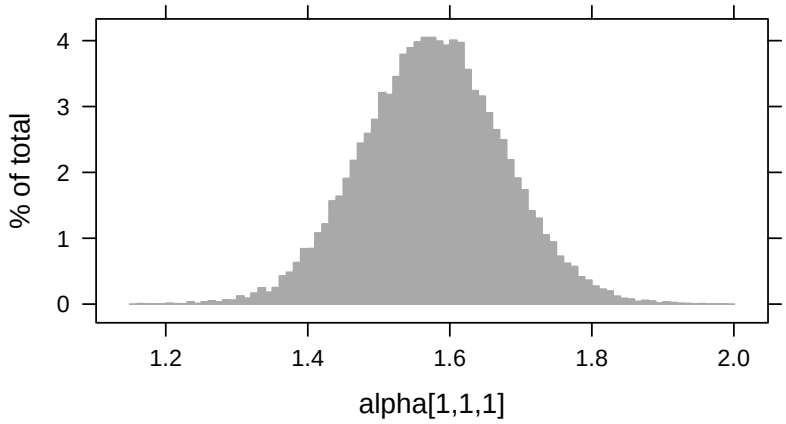


Fig. 6.1: Samples from the posterior for α_{i}.

One practical advantage of a 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.

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

```
## alpha[1,1,1] alpha[2,1,1] alpha[3,1,1]
##         1.65         0.35         0.83
## theta[1,1,1] theta[2,2,1] theta[3,3,1]
##         0.69         0.38         0.48
```

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. We can feed each set of model parameter values 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 require that this future dataset has the sample size of the new dataset in order to get a prior predictive distribution with data sets comparable to the new data set. 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 and predict a dataset for each of those samples.

Within one function, `ppc.step1`, the R-package `Replication` can (1) obtain the posterior distribution for the original data accompanied by traceplots and a summary of the results for verification, (2) draw samples from the posterior distribution, and (3) simulate future data. The function `ppc.step1` 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 `blavaan` 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. 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))
```

The 5,000 datasets that we obtain as a result, 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  1.1462610 1.612134 2.223643
## 2  2.2310011 1.669518 2.172725
## 3 -0.5843558 3.381702 3.759951
## 4  0.6607782 1.729814 2.525175
## 5  2.4944977 2.352052 3.325300
## 6  1.5036758 1.376774 3.476710
```

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 of 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.

6.4 The Replication Hypothesis H_0

The findings of the original study can be summarized in an informative hypothesis H_0 (Hoijtink, 2012; Silvapulle and 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 and 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 replication 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 replication hypothesis H_0 is determined, it 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 replication 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.

6.5 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 et al. (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.

6.6 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 6.3. 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 6.6). 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 is not aimed at 'proving' 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 6.6, 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 6.2 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 6.2. 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).

Generally, the prior predictive p-value tells us how extreme the new study scores compared to 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

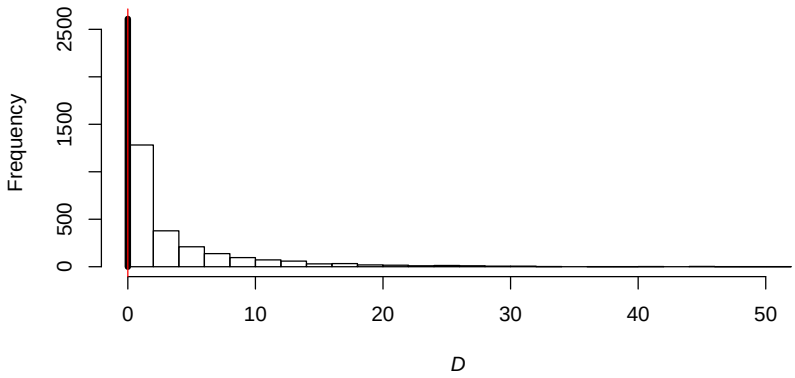


Fig. 6.2: Histogram of predicted D for the replication of Achterberg et al. (2017) with the observed D for Achterberg et al. (2018) indicated by the red line

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 the 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 can 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.

6.7 An Example of a Multilevel Longitudinal Growth Curve Model

Bakker et al. (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 6.3. The top of Figure 6.3 shows the time level with three repeated measurements of intrusion: int0, int3, int12 and int18. The model contains three latent growth factors: (1) i, the intercept of intrusion at the first measurement, (2) s, the linear growth rate per month at the first measurement,

and (3) q, the quadratic factor. 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 et al. (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.

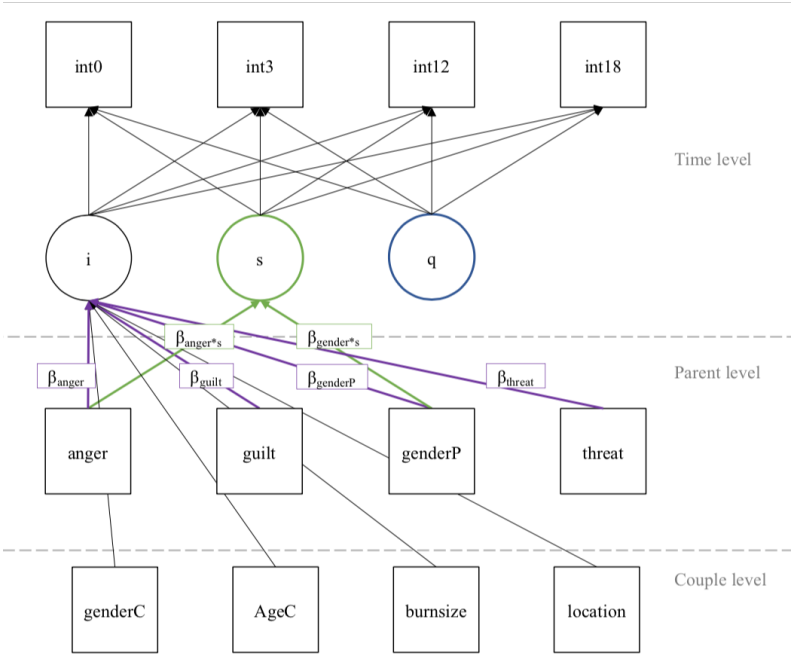


Fig. 6.3: Multilevel model as evaluated in Bakker et al. 2013.

6.7.1 The Prior Predictive Distribution

First, the multilevel model was rewritten at the first level for wide format data (see Supplementary R-code at `https://osf.io/as7kz`), 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 replication hypothesis of interest H_0 with which we could compare the predicted datasets to the observed new dataset.

6.7.2 The Replication Hypothesis H_0

With respect to intrusion, Bakker et al. (2013) drew conclusions about the eight bold and colored parameters in Figure 6.3:

> "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 the difference in child age is not a reason to expect different values for the latent time variables and regression parameters of interest. Hence, H_0: s < -0.63, q > 0.02, $\beta_{\text{genderP}} > 4.84$, $\beta_{\text{guilt}} > 0.56$, $\beta_{\text{anger}} > 1.30$, $\beta_{\text{threat}} > 2.07$, $\beta_{\text{genderP*s}} < -0.08$, $\beta_{\text{anger*s}} < -0.06$. Again, we identify the `plables` and include them in an object $H0$ (see Supplementary Materials at `https://osf.io/as7kz` for an automatized procedure).

6.7.3 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 = −0.56, q = 0.02, $\beta_{\text{genderP}} = 5.41$, $\beta_{\text{guilt}} = 1.19$, $\beta_{\text{anger}} = 0.58$, $\beta_{\text{threat}} = 1.39$, $\beta_{\text{genderP*s}} = -0.18$, $\beta_{\text{anger*s}} = 0.00$. 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 step 1 stored in `step1.B.M`, (2) the new data `y.r`, and (3) the replication 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) $D = 10.89$, and the prior predictive p-value is 0.013 (See also Figure 6.4). The new data by Egberts et al. (2017) scored in the extreme 1.3% of the predicted data with respect to the replication hypothesis. Hence, we reject the replication of H_0 by Egberts et al. (2017).

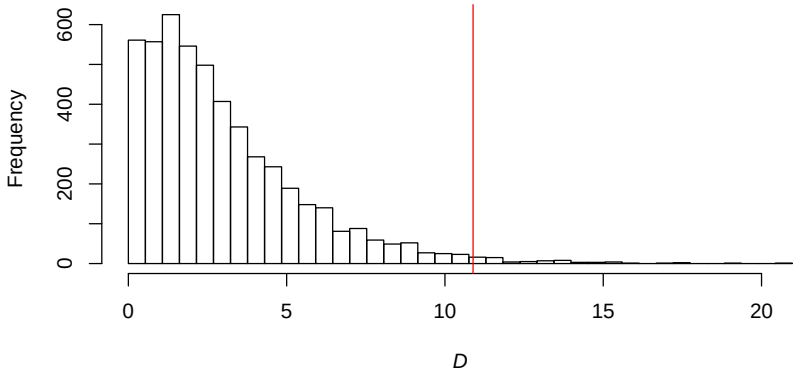


Fig. 6.4: Histogram of predicted D for the replication of Bakker et al. (2013) with the observed D for Egberts et al. (2017) indicated by the red line.

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, we used 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 $\boldsymbol{y}_r$.

6.7.4 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 and 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,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, p. 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 in this example.

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 replicates 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 6.5 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 6.6.

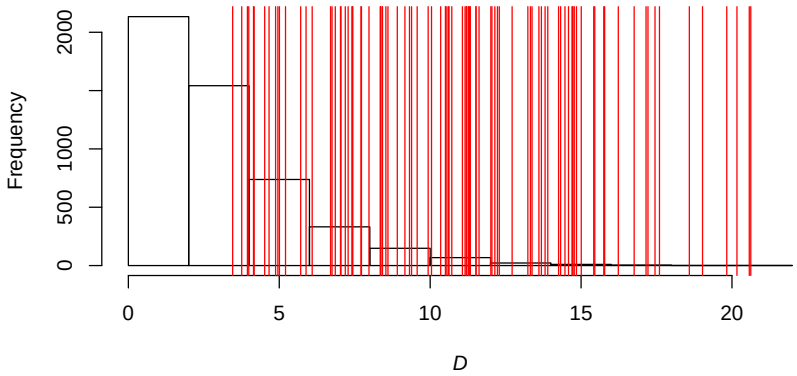


Fig. 6.5: Histogram of D for the predicted data with scores for the imputed new datasets in red.

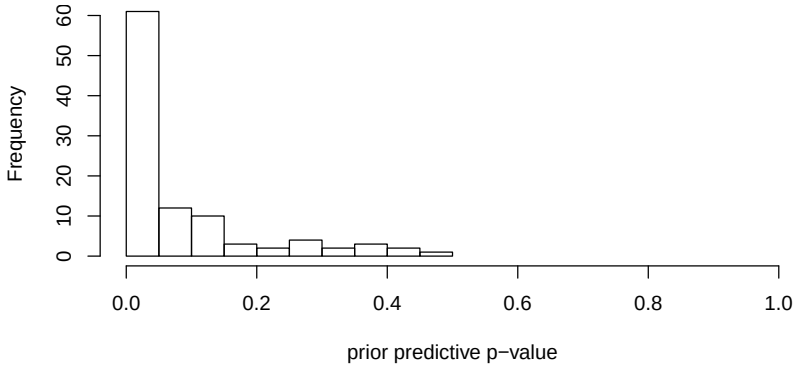


Fig. 6.6: Histogram of prior predictive p-values for the imputed datasets.

The computed p-values over the imputations vary from 0.00 to 0.63. 59.0 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.

6.8 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.

B

Appendices

B.1 Syntax (b)lavaan Model for Achterberg et al. (2017)

```
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.

B.2 Prior Specifications Bayesian Analyses

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 B.1.

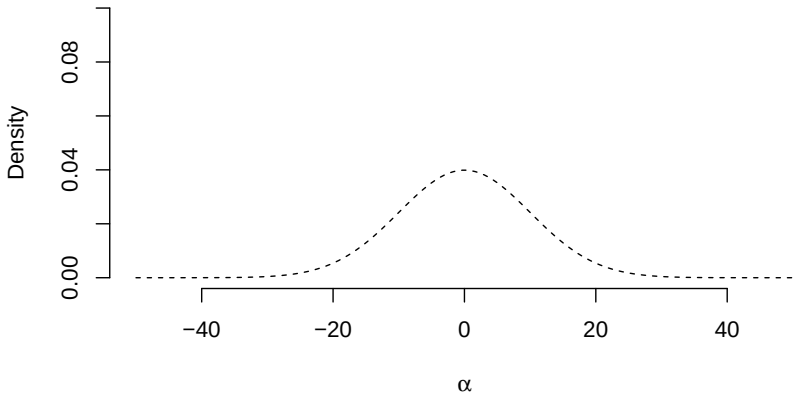


Fig. B.1: Prior distribution for $\boldsymbol{\alpha}$.

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 B.2.

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

```
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).

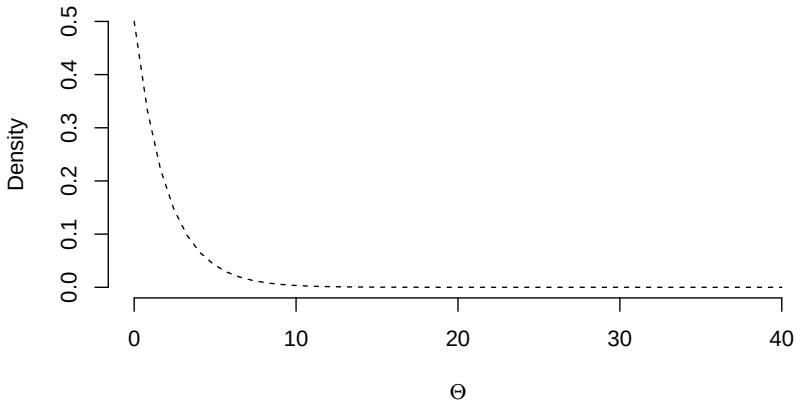


Fig. B.2: Prior distribution for $\boldsymbol{\Theta}$.

7

Testing Replication with Small Samples: Applications to ANOVA

Summary. Findings based on small samples can offer important insights, but the original small sample findings should be replicated before strong conclusions can be drawn. This chapter describes some of the difficulties that arise when attempting to replicate findings from small sample research. We present four common replication research questions: 1) whether the new effect size is similar to the original effect size; 2) whether the new effect size differs from the original effect size; 3) whether the conclusions based on new results differ from the original conclusions; and 4) what the effect size is in the population. Appropriate evaluation methods are discussed for each of these research questions: the replication Bayes factors, confidence intervals, methods based on prediction intervals, and bias-corrected meta-analysis. Each method is illustrated for the replication of an ANOVA and associated post-hoc *t*-tests. Annotated R-code for all analyses is provided with the chapter.

7.1 Introduction

Concerns about the replicability of studies were expressed as early as in 1979 by Rosenthal, who believed that future insights would solve this problem. The field of psychological science, however, is still struggling to establish replicability, as was clearly shown with the Reproducibility Project Psychology (RPP; Open Science Collaboration, 2015). Increased awareness of the fuzziness of results obtained using small samples is an important step towards improving this situation (Lindsay, 2015). Results obtained

This chapter is accepted as Zondervan-Zwijnenburg, M.A.J., & Rijshouwer, C.D.N. Testing Replication with Small Samples: Applications to ANOVA. In: R. van de Schoot, M. Miočević (Eds.), *Small Sample Size Solutions: A guide for applied researchers and practitioners*. Routledge.

Author contributions: MZ and DR were involved in the initial research design. DR drafted the initial chapter evaluating only research question 3. MZ wrote the final chapter evaluating four research questions with different methods. MZ conducted the analyses. DR provided additional feedback.

with smaller samples are less likely to be replicated than those obtained with larger samples (Cohen, 1962).

One of the difficulties in replicating small sample research is that small samples are particularly sensitive to 'researcher degrees of freedom': decisions that researchers make in the design and analysis of the data (Simmons et al., 2011). For example, researchers decide to combine categories, exclude scores, add comparisons, add covariates, transform measures, etc. Unfortunately, modifications are more common if results do not support the hypothesis. For example, the impact of an extreme score will more often be detected and adjusted if it causes a non-significant result as compared to a significant result. With small samples, these decisions can easily affect the significance of results, leading to inflated false positive rates (Simmons et al., 2011).

Another issue is publication bias: studies with statistically significant results are published more often than studies with non-significant results. Small sample studies are often underpowered, leading to non-significant results and hence a reduced chance to be published. On the other hand, small studies that do find significant effects appear impressive and are more likely to be published.

Thus, researcher degrees of freedom and publication bias can lead to overestimation of effects and an inflated false positive rate in the literature (Simmons et al., 2011). Small sample findings therefore can easily be spurious, meaning that their replication is of great importance.

Different replication research questions require different methods. Here, we distinguish four main research questions that can be investigated if a new study is conducted to replicate an original study:

1. Is the new effect size similar to the original effect size?
2. Is the new effect size different from the original effect size?
3. Are the conclusions based on new results different from the original conclusions?
4. What is the effect size in the population?

Note that questions one and two differ in where the burden of proof lies. Question one looks to *provide support* for the equality of effect sizes, whereas question two is aimed at *falsifying* the claim of equality of effect sizes in favor of a conclusion that the effect size was not replicated.

For all four replication research questions we recommend statistical methods and apply them to an empirical example. Note that Anderson and Maxwell (2016) also documented replication research questions and associated methods, although not specifically for small samples. In the current chapter, we adopt several suggestions from Anderson and Maxwell (2016) and add more recent methods. R-code (R Core Team, 2017) for reproducing all chapter results is provided as Supplementary Material at `https://osf.io/x3ua2`. We demonstrate the four replication research methods for the replication of Henderson et al. (2008) by Lane and Gazerian (2016). First, we introduce the original study by Henderson et al. (2008) and its replication by Lane and Gazerian (2016). This is followed by a discussion of the four replication research questions and their associated methods.

7.1.1 Original Study and its Replication

Henderson et al. (2008) conducted a series of experiments showing that people who planned the implementation of a chosen goal (i.e., people with an "implemental mind-set") have stronger attitudes, even towards topics unrelated to their actions. Experiment 5 is the one that was replicated by Lane and Gazerian (2016). It is designed to demonstrate that a focus on information that supports the previously made decision is the reason that attitude strength increases with an implemental mind-set. The experiment included three conditions with 46 participants in total. The first condition was a neutral condition in which participants described things they do on a typical day. The second condition was an implemental one-sided focus condition. Participants in this condition chose a romantic topic to write about and wrote down three reasons for that choice. The third condition was the implemental two-sided focus condition in which participants made their choice and wrote down three reasons for and three reasons against this choice. Afterwards, participants in all conditions answered three questions rating their attitude ambivalence with respect to the issue of making public a list with names of convicted sex offenders (e.g., "I have strong mixed emotions both for and against making the list of convicted sex offenders available to the general public rather than just the police").

The descriptive statistics of the data for the experiment by Henderson et al. (2008) are provided in Table 7.1. The effect of the conditions on attitude ambivalence was significant as Henderson et al. (2008) report: $F(2, 43) = 3.36$, $p = .044$, $\eta^2 = 0.13$, $\omega^2 = 0.09$, $r = 0.26$. We have added the effect size ω^2, because it is less biased than η^2 for small samples (Okada, 2013). Furthermore, we also computed the effect size r as used in the RPP as an additional effect size measure (see Appendix 3 at `https://osf.io/z7aux`). Assuming that all predictors (i.e., the dummy condition variables) contributed equally to the explained variance, r^2 is the explained variance per predictor, and r is the correlation coefficient per predictor.

Post-hoc comparisons revealed that the implemental mind-set one-sided group demonstrated significantly lower amounts of ambivalence compared to the implemental mind-set two-sided group, $t(28) = 2.45$, $p = .021$, Cohen's $d = .93$, Hedges' $g = .50$. For the t-test, we added Hedges' g to correct for an upwards bias that Cohen's d shows with small samples. Hedges' g is obtained by multiplying Cohen's d by the correction factor $(1 - \frac{3}{4df-1})$ (Hedges, 1981). The mean of the neutral mind-set group was in the middle, but it was not significantly higher or lower than the means of other conditions (see descriptive statistics in Table 7.1). Henderson et al. (2008) write: "Critically, the findings showed that it was the evaluatively one-sided analysis of information, rather than simply the act of deciding itself, that fostered a spillover of decreased ambivalence ... " (p. 406-407).

Lane and Gazerian (2016) replicated the experiment with 70 participants, but found no significant effect of condition on ambivalence, $F(2, 67) = 1.70$, $p = .191$, $\eta^2 = 0.05$, $\omega^2 = .02$, $r = 0.16$ (see also the descriptive statistics in Table 7.1). The post-hoc difference test between the one- and two-sided implemental mind-set groups was not significant, $t(44) = 1.24$, $p = .222$, Cohen's $d = .36$, Hedges' $g = .25$. Based on

the lack of significance in the new study, Lane and Gazerian conclude that the effect may not replicate.

7.2 Four Replication Methods

Evaluating the significance (and direction) of the effect in the new study and using it as a measure for replication, as was a main method of Lane and Gazerian (2016), is called 'vote-counting'. Vote-counting, however, does not take into account the magnitude of the differences between effect-sizes (Asendorpf et al., 2013; Simonsohn, 2015); it is not a statistical test of replication (Anderson and Maxwell, 2016; Verhagen and Wagenmakers, 2014); and it leads to misleading conclusions in underpowered replication studies (Asendorpf et al., 2013; Simonsohn, 2015). Thus, vote-counting is a poor method to assess replication. In the following, we discuss four alternative replication research questions and methods.

	Neutral		One-sided implemental		Two-sided implemental	
	n	M (SD)	n	M (SD)	n	M (SD)
Original	16	1.23 (1.64)	15	0.16 (1.85)	15	1.82 (1.86)
New	24	-0.38 (1.44)	23	-0.14 (1.66)	23	0.39 (1.25)

Table 7.1: Descriptive Statistics for Confirmatory Information Processing from the Original Study: Henderson et al. (2008), and the New Study: Lane and Gazerian (2016)

7.2.1 Question 1. Is the New Effect Size Similar to the Original Effect Size?

A frequentist approach to this replication research question is the equivalence test (e.g., Walker and Nowacki, 2011). This test requires the researcher to specify a region of equivalence for the difference between the original and new effect size. If the confidence interval of the difference between effects falls entirely within this region, the effect sizes are considered equivalent. However, it is difficult to set a region of equivalence that is reasonably limited while at the same time the confidence interval for the difference between effects has a chance to entirely fit within the interval. Therefore, we do not elaborate on the equivalence test and focus instead on Bayesian approaches.

To evaluate whether the new effect size is similar to the original effect size, we can compute a Bayes factor (BF; Jeffreys, 1961). A BF expresses the shift in belief, relative to our prior belief, after observing the data for two competing hypotheses. A BF of 1 is undecided. BFs smaller than one indicate preference for the null hypothesis, whereas BFs larger than one favor the alternative hypothesis. The two competing hypotheses in the BF can be operationalized in many ways, but in the replication setting, one of the

evaluated hypotheses is often the null effect (i.e., the effect size is zero). To evaluate the current research question, a proper alternative hypothesis is that the effect in the new study is similar to the effect in the original study Harms (2018a); Ly et al. (2018); Verhagen and Wagenmakers (2014). In this case, the BF evaluates whether the new study is closer to a null effect, or closer to the original effect, where the original effect forms the prior distribution in the BF for the new effect. Verhagen and Wagenmakers (2014) developed this BF for the t-test. Harms (2018a) extended the Replication BF to the ANOVA F-test and developed the `ReplicationBF` R-package (Harms, 2018b) to compute it based on the sample sizes and test statistics of the original and new study. For the ANOVA by Henderson et al. (2008) replicated by Lane and Gazerian (2016), we obtain a Replication BF of 0.42, which means that the evidence for the null hypothesis of no effect is 2.40 (i.e., 1 / 0.42) times stronger than the evidence for the alternative hypothesis that the effect is similar to that in the original study. See Figure 7.1 for a visualization by the `ReplicationBF` package. The R-package also includes the Replication BF for t-tests as proposed by Verhagen and Wagenmakers (2014). For the post-hoc t-test we find a Replication BF of 0.722, which is again in favor of a null effect. Thus, the Replication BF does not support replication of the omnibus ANOVA effect, nor does it support the replication of the post-hoc result that the one-sided mind-set group scores lower on ambivalence than the two-sided mind-set group.

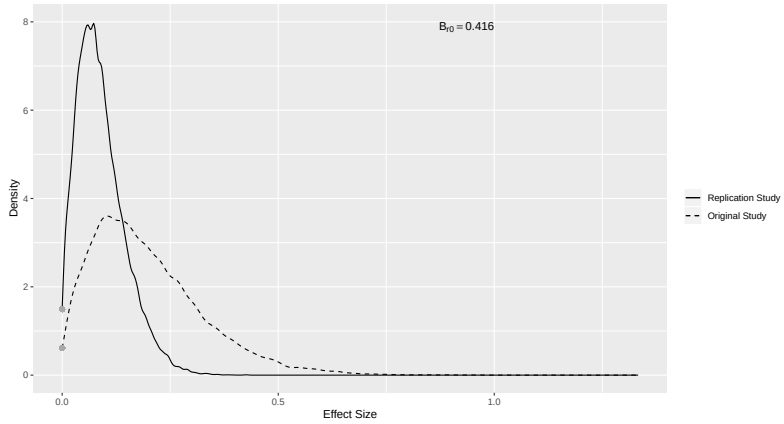


Fig. 7.1: The Replication BF by Harms (2018b). The original study is the prior for the effect size and the replication study is the posterior based on that prior and the new study. The ratio of the two distributions at 0 on the x-axis is the Replication BF.

Ly et al. (2018) provided a simple calculation to obtain the Replication BF by Verhagen and Wagenmakers (2014) for all models for which a BF can be obtained:

We report BFs up to two decimal places, but use all available information for calculations

Evidence Updating (EU) Replication BF $= \frac{\text{BF combined data}}{\text{BF original data}}$. This calculation assumes, however, that the data are exchangeable (see Chapter 2 for a discussion on exchangeability). If the original and new study are not based on the same population, the combined data may demonstrate artificially inflated variances due to different means and standard deviations. To minimize the impact of non-exchangeable datasets, Ly et al. (2018) suggest to transform the data. Here, the grand mean in Henderson et al. (2008) is actually 1.03 points higher than the grand mean in Lane and Gazerian (2016). To address this issue, we converted the responses to z-scores.

To compute the BFs for the combined and original data, we can use the point-and-click software JASP (JASP Team, 2018) or the BayesFactor package (Morey, 2018) in `R`. For both software packages, the BF for the combined data is 1.50, and the BF for the original data is 1.59. Hence, the EU Replication BF = 1.50 / 1.59 = 0.94, which favors the ANOVA null hypothesis that the effect is zero. For the post-hoc analysis with the alternative hypothesis that the one-sided mind-set group scores lower than the two-sided mind-set group, the BF for the combined data is 6.66 (see Figure 7.2 for the accompanying `JASP` plot), and the BF for the original data is 5.809. Hence, the EU Replication BF = 6.66 / 5.81 = 1.15 for the replication of the original effect. Thus, the EU Replication BF is ambiguous about the replication of the omnibus ANOVA effect (i.e., BF = 0.94), nor does it provide strong support for the replication of the post-hoc result.

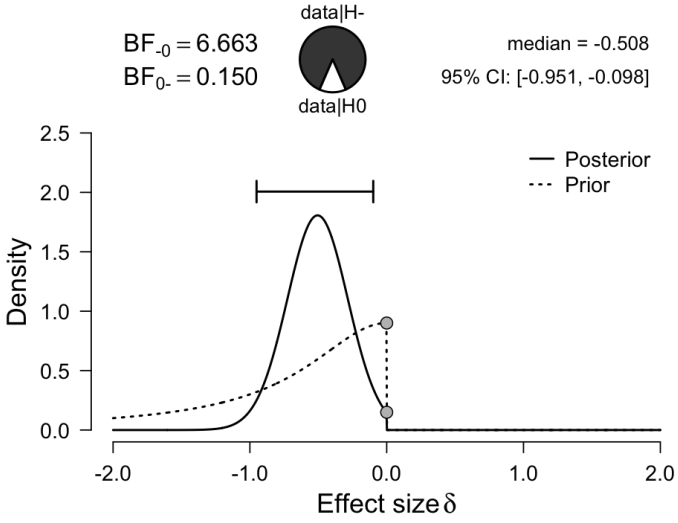


Fig. 7.2: BF with default prior settings in the combined data for the one-sided t-test. The ratio of the two distributions at 0 on the x-axis is the BF.

Note that the BFs according to the method presented in Ly et al. (2018) are higher than those calculated by the `ReplicationBF` package by Harms (2018b), even though both are extensions of Verhagen and Wagenmakers (2014). Harms (2018a) and Ly et al. (2018) discuss several differences between both approaches: (1) both methods use different priors (i.e., uniform in `ReplicationBF` R-package, Cauchy in `JASP` and the `BayesFactor` R-package, (2) the EU Replication BF assumes exchangeability to compute the BF for the combined data, and (3) for ANOVA models the BF computed in the `ReplicationBF` package is based on the sample size and test statistics, whereas `JASP` and the `BayesFactor` package use a more elaborate model that involves the full dataset(s). `JASP` currently also has a Summary Statistics module for t-tests, regression analyses and analyses of frequencies. Whenever possible, we recommend applying both methods to obtain a more robust evaluation of replication.

7.2.2 Question 2. Is the New Effect Size Different from the Original Effect Size?

To test whether the new effect size is different from the effect size in the original study, we would preferably compute a confidence interval for the difference in effect sizes. The literature does not provide such an interval for η^2 or ω^2. However, with an iterative procedure based on descriptive statistics we can obtain separate confidence intervals for ω^2 in the original and new study (Steiger, 2004). Let us denote the original study with subscript o, and the new study with subscript n. For the original study $\omega_o^2 = .09$, 95% CI $[.00, .30]$ (see Supplementary Materials for all calculations). For the new study $\omega_n^2 = .02$, 95% CI $[.00, .22]$. With these confidence intervals, we can calculate a confidence interval for the difference between both effect sizes, $\Delta\omega^2$ by applying the modified asymmetric method introduced by Zou (2007) for correlations and squared correlations. This method takes into account that some effect sizes have asymmetric distributions or cannot take on negative values (such as ω^2). $\Delta\omega^2 = .07$, 95% CI $[-.15, .29]$. Since zero is in the confidence interval of the difference between the effect sizes, we do not reject the hypothesis that the effect sizes are equal, and thus, we retain the hypothesis that the new effect replicates the original one.

For the post-hoc difference between the one-sided and two-sided implemental conditions we can compute the 95% confidence interval for standardized mean differences (i.e., Cohen's $d_o = .93$ and Cohen's $d_n = .36$) as given in Bonett (2009) and included in the Supplementary Materials. The difference between Cohen's d for both studies is 0.57, 95% CI $[-.0.96, 2.10]$. Since zero lies in the confidence interval, we do not reject replication of the original effect size.

Alternatively, Patil et al. (2016) describe how non-replication of an effect size can be tested with a prediction interval. A 95% prediction interval aims to include the (effect size) estimate in the next study for 95% of the replications. Patil et al. (2016) (see Supplementary Materials) apply this method on r as calculated by the RPP. Following their methods, we find that the prediction interval for $r_o = .26$ ranges from -0.12 to 0.57. The estimate for the new study, $r_n = 0.16$, lies within the interval of estimates that are expected given replication (i.e., -0.12 to 0.57). Hence, we do

not reject replication of the original effect size. Note that Patil et al. (2016) apply their method on r, which is considered problematic when r is based on more than two groups (see, for example, Appendix 3 at `https://osf.io/z7aux`). The post-hoc t-test value of r_o is .42, with a prediction interval ranging from −0.03 to 0.73. For the new study, $r_n = .18$. Again, the correlation estimate for the new study lies within the prediction interval, and we do not reject the hypothesis that the original effect has been replicated.

The confidence intervals for the difference between effect sizes and the prediction intervals in this example can be considered to be quite wide. If the study results are uncertain (i.e., based on small samples), the associated confidence and prediction intervals will less often reject replication of the original effect size. However, especially with small studies, a failure to reject replication does not necessarily imply replication, but rather a lack of power, which suggests that the above methods may be inadequate for small samples.

7.2.3 Question 3. Are the Conclusions based on New Results Different from the Original Conclusions?

In contrast to the first two replication research questions which concerned effect sizes, the current question concerns conclusions. The prior predictive p-value can be used to answer this question (Box, 1980; Zondervan-Zwijnenburg et al., 2019). The calculation of the prior predictive p-value starts with the simulation of datasets from the predictive distribution (with the sample size used in the new study) that are to be expected, given the original results. Subsequently, the new observed data from the replication attempt are compared to the predicted data with respect to a replication hypothesis. The replication hypothesis includes the conclusions of the original study in an informative hypothesis (Hoijtink, 2012) This hypothesis can include the ordering of parameters (e.g., $\mu_1 > \mu_2$), the sign of parameters (e.g., $\mu_1 > 0$, $\mu_2 < 0$), or the exact value of parameters (e.g., $\mu_1 = 3$, $\mu_2 = -2$). Any combination of constraints is possible. The deviation from the hypothesis for each of the predicted datasets and for the new dataset is expressed in the statistic that we call $\bar{F}$. With $\alpha = .05$, replication of the study's conclusions is rejected if the misfit with the replication hypothesis in the new study is equal to or higher than in the extreme 5% of the predicted data. All computations can be conducted in an online interactive application presented at `osf.io/6h8x3` or with the `ANOVAreplication` R-package (Zondervan-Zwijnenburg, 2018).

The results and conclusion of Henderson et al. (2008) lead to the following replication hypothesis: $\mu_{\text{One-sided implemental}} < (\mu_{\text{Two-sided implemental}}, \mu_{\text{Neutral}})$, Cohen's $d_{\text{One-sided implemental,Two-sided implemental}} > .8$. If we run the test, we find that the prior predictive p-value = .130. Hence, we do not reject replication of the original study's conclusions. Figure 7.3 shows the statistic $\bar{F}$ for each of the predicted datasets and the replication by Lane and Gazerian (2016). Note that we do not have to run a post-hoc analysis with this method, because the conclusion for the post-hoc contrast was incorporated in the replication hypothesis with "Cohen's $d_{\text{One-sided implemental,Two-sided implemental}} > .8$".

For the prior predictive p-value, an original study with large standard errors (e.g., due to a small sample) leads to a wide variety of predicted datasets, thus making it hard to reject replication of the original study conclusions. With the `ANOVAreplication` `R`-package we can calculate the power to reject replication when all means would be equal in the new study. Here, the statistical power was only .57. The sample size in the new study also affects power to reject replication of the original study conclusions. When we calculate the required sample size to obtain sufficient power, we find that the statistical power stagnates around .63, even for samples of 200 per group, due to large standard errors in the original study.

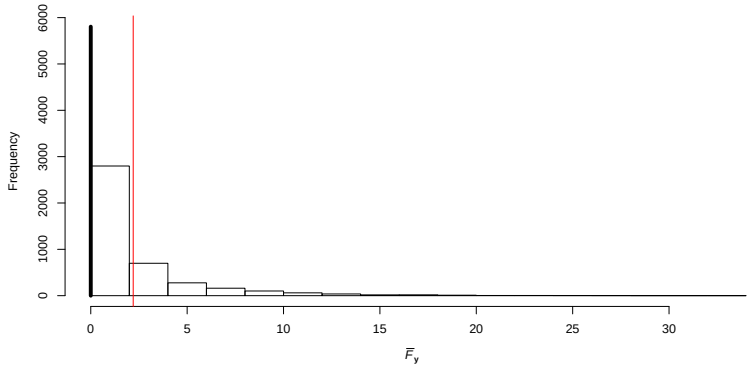


Fig. 7.3: Prior predictive p-value. The histogram concerns $\bar{F}$-scores for each of the 10,000 predicted datasets with respect to the replication hypothesis. The thick black line represents the 5,805 predicted datasets that had an $\bar{F}$-score of exactly 0 and were perfectly in line with the replication hypothesis. The red line indicates the $\bar{F}$-score of 2.20 for the new study. The $\bar{F}$-score for the new data is positioned in the extreme 13.0% of the predicted data (prior predictive $p = .130$).

7.2.4 Question 4. What Is the Effect Size in the Population?

At the end of the day, most researchers are concerned with the effect in the population. To determine the population effect based on an original and new study, numerous meta-analytic procedures have been proposed. For close replications, the fixed-effect meta-analysis can be used, which assumes that there is one underlying population from which both studies are random samples. Consequently, there is only one underlying true effect size. However, the standard fixed-effect meta-analysis does not take publication bias into account. As a result, standard fixed-effect meta-analyses overestimate effect sizes.

The frequentist hybrid meta-analysis (Van Aert and Van Assen, 2017b) and the Bayesian snapshot hybrid method (Van Aert and Van Assen, 2017a) are two meta-analytic methods developed for situations with a single replication effort that take into account the significance of the original study (which could be caused by publication bias). Both methods are part of the `puniform` R-package (Van Aert, 2018), and available as online interactive applications. The frequentist hybrid meta-analysis results in a corrected meta-analytic effect size and its associated confidence interval and p-value. The output also includes the results of a standard fixed-effect meta-analysis for comparison. The Bayesian snapshot hybrid method quantifies the relative support, given the original and replication study, for four effect size categories: zero, small, medium, and large. Currently, both methods can be used for correlations and t-tests. However, the correlation for the original ANOVA as computed by the RPP cannot be used for the meta-analytic methods, because its standard error cannot be computed for more than two groups.

For the post-hoc t-test results of Henderson et al. (2008) and Lane and Gazerian (2016), the bias-corrected Hedges' g is .37, 95% CI $[-.48, .94]$, $p = .232$. Thus, we cannot reject the hypothesis that the effect in the population is zero. The standard (uncorrected) fixed-effect meta-analytic estimate was .60, 95% CI [0.12,1.07], p=.014. Whereas the fixed-effect meta-analytic effect sizes was significant at $\alpha = .05$, the hybrid meta-analysis effect size is lower and has a wider 95% confidence interval. The snapshot hybrid method with equal prior probabilities for the four effect size categories indicated that a small effect size received the highest support (37.8%), followed by no effect size (30.2%), a medium effect size (25.5%), and a large effect size (6.6%).

Besides meta-analyses that take significance of the original study into account, we can also calculate the Bayes factor for an effect versus no effect, based on the scaled combined data using `JASP`. The Bayes factor in favor of an ANOVA effect is 1.50. The Bayes factor in favor of a post-hoc t-test effect is 6.66. Hence, the evidence in the combined data is positive with respect to the existence of an effect. Note that this combined analysis does not correct for publication bias and assumes exchangeability. Alternatively, Etz and Vandekerckhove (2016) developed a Bayes factor for t-tests, univariate F-tests (i.e., not more than two groups), and univariate regression analyses that takes into account publication bias, but unfortunately this Bayes factor has only been developed for the `Matlab` software package, which is mainly used by engineers, mathematicians, and economists.

7.3 Discussion

In this chapter, we presented replication research questions and associated statistical techniques. In the example we used, the replication BFs pointed mostly towards a null-effect instead of a replication of the original effect; the confidence intervals around the difference between effect sizes indicated that the difference between the original and new study may be zero, but they had low power; the prior predictive p-value could also not reject replication of the original study's conclusions; and meta-analyses

indicated that the population effect is small, anecdotal, or not significantly different from zero.

We also discussed how the different methods perform with small samples. Bayes factors and the Bayesian snapshot meta-analysis have the advantage over null hypothesis significance testing (NHST) methods (e.g., confidence intervals and the prior predictive p-value) that they cannot be underpowered. The evidence by the BF may not be overwhelming, but at least it indicates the relative plausibility of one hypothesis over the other after observing the data. NHST methods, on the other hand, often result in non-significant findings with small samples, and it remains unclear whether the (non)replication effect was absent, or whether the analysis was underpowered.

An advantage of the prior predictive p-value is that it allows the user to test the replication of the original study's conclusions summarized in a replication hypothesis. This hypothesis can include multiple parameters, and it can convey information on their size and ordering. In the ANOVA setting, the effect size (e.g., η^2) does not provide information about the direction of the effect. Hence, it is useful to evaluate an informative replication hypothesis that specifies the ordering of group means.

The preferred method to test replication depends on the replication research question at hand. Furthermore, given a replication research question, it can be insightful to apply multiple methods to test replication (Harms, 2018a). Testing replication yields more meaningful results with larger sample sizes, and this holds for all methods described in this chapter. Testing replication of small sample research is challenging, but since small samples are more susceptible to researcher degrees of freedom, it is of utmost importance to critically evaluate small sample results with replication studies.

Part III

Cross-Validating and Synthesizing Information From Multiple Cohort Studies

8

Parental Age and Offspring Childhood Mental Health: A Multi-Cohort, Population-Based Investigation

Summary. To examine the contributions of maternal and paternal age on offspring externalizing and internalizing problems, this study analyzed problem behaviors at age 10-12 years from four Dutch population-based cohorts (N = 32,892) by a multiple informant design. Bayesian evidence synthesis was used to combine results across cohorts with 50% of the data analyzed for discovery and 50% for confirmation. There was evidence of a robust negative linear relation between parental age and externalizing problems as reported by parents. In teacher-reports, this relation was largely explained by parental socio-economic status. Parental age had limited to no association with internalizing problems. Thus, in this large population-based study, either a beneficial or no effect of advanced parenthood on child problem behavior was observed.

Since 1995, the mean maternal age at first birth has increased at a rate of 0.10 years per year in OECD countries, and in 2017 exceeded 30 years in the vast majority of these countries (Organisation for Economic Co-operation and Development, 2017).

This chapter is published as Zondervan-Zwijnenburg, M.A.J.*, Veldkamp, S.A.M.*, Neumann, A., Barzeva, S.A., Nelemans, S.A., Van Beijsterveldt, C.E.M. Branje, S., Meeus, W.H.J., Hillegers, M.H.J., Tiemeier, H., Hoijtink, H.J.A., Oldehinkel, A.J., & Boomsma, D.I. (2019). Parental Age and Offspring Childhood Mental Health: A Multi-Cohort, Population-Based Investigation. *Child Development*. doi: 10.1111/cdev.13267

* These authors contributed equally.

Author contributions: HH, DB, and TO initiated the project. MZ and SV managed the project. MZ, SV and HH designed the data analysis protocol, supervised its execution, and summarized the outcomes. SV, MZ, HH, TO, and DB wrote the manuscript. SV, AN, SBa, SN, and MZ contributed to the exploratory analysis design and analyzed data. AN and HT contributed to Gen-R data-collection. TvB and DB, contributed to the NTR data collection. TO contributed to the TRAILS data collection. SN, SBr, and WM contributed to the Radar-Y data collection. All authors read, reviewed, revised, and approved the manuscript.

Only in Mexico was the mean age of women at childbirth lower than 28 years, and only in eight countries was it between 28 and 30 years of age. Women's reproductive years generally range from about 15 to 45 years (Te Velde, 2002). Within this wide age range some periods are generally considered more suitable to have children than others, but which parental reproductive ages are optimal for offspring physical and mental health has been a matter of debate ever since individuals have engaged in active birth control. Whereas having children at an advanced age was quite common historically, when families tended to be larger (e.g. Desjardins et al., 1994), the current trend to delay childbearing has given rise to public health concerns.

8.0.1 Concerns Regarding Delayed Childbearing

Concerns regarding delayed childbearing are understandable, as a large number of research reports highlight that increased maternal age at childbirth is associated with several adverse consequences, ranging from physical problems, such as increased BMI, blood pressure and height (Carslake et al., 2017) to psychiatric conditions, such as autism (Lee and McGrath, 2015; Sandin et al., 2012), bipolar disorder (Menezes et al., 2010), symptoms of depression, anxiety and stress (Tearne et al., 2016), and poor social functioning (Weiser et al., 2008). More recently, increased paternal age at birth has also been associated with adverse child outcomes, such as stillbirth and cleft palate (see for a review Nybo Andersen and Urhoj, 2017). In over 40 million live births between 2007 and 2016, having an older father increased the risk of low birthweight, apgar score, and premature birth (Khandwala et al., 2018). A study of the Danish population, which included 2.8 million persons, found that older fathers are at risk of having offspring with intellectual disabilities, autism spectrum disorders and schizophrenia (McGrath et al., 2014; de Kluiver et al., 2017).

Several, not mutually exclusive, mechanisms have been proposed to explain the increased physical and mental health risks in offspring of older parents. First, age-related deterioration of the functioning of women's reproductive organs, such as DNA damage in germ cells, and worse quality of oocytes and placenta, can increase the risk of obstetric and perinatal complications (Myrskylä and Fenelon, 2012). Second, male germline cells undergo cell replication cycles repeatedly during aging, with de novo point mutations accumulating over time (e.g., Jónsson et al., 2017) and the number of de novo mutations in the newborn increasing with higher age of the father at the time of conception (Kong et al., 2012; Francioli et al., 2015). Although weaker than with paternal age, de novo mutations in offspring correlate with maternal age as well (Goldmann et al., 2018; Wong et al., 2016). Third, genomic regions in the male germline may become less methylated with increasing age (Jenkins et al., 2014) and alter the expression of health-related genes. Fourth, age effects can be due to selection, with older parents differing from younger ones in characteristics that are relevant for developmental outcomes in their offspring, such as poor social skills. The influence of selection effects can be exacerbated by assortative mating (Gratten et al., 2016). Fifth, being the child of older parents carries the risk of having to cope with parental frailty or losing a parent at a relatively young age (Myrskylä and Fenelon, 2012), and

the stress evoked by these experiences may trigger health problems. Most of these mechanisms involve consequences of biological ageing. Parenthood at an advanced age is disadvantageous from a biological perspective; except for very young, physiologically immature mothers, younger parents are in a better physical condition.

8.0.2 Possible Benefits of Delayed Childbearing

Whereas the effects of older parental age on children's physical health and psychiatric disorders tend to be predominantly negative, the effects of older parental age on mental health problems with a stronger psychosocial component, such as externalizing and internalizing problems, tend to be more inconsistent. An indication that the negative consequences of high parental age may stretch beyond clinical diagnosis is provided by Tearne et al. (2015, 2016), who found that high maternal age predicted symptoms of depression, anxiety and stress in daughters, and by Janecka et al. (2017a) who reported a negative association between advanced paternal age and social development. In contrast, in several population-based studies, offspring of older parents, particularly of older mothers, perform better at school and work, score higher on intelligence tests, report better health and higher well-being, use fewer drugs, and have fewer behavioral and emotional problems than offspring of younger parents (e.g., Carslake et al., 2017; McGrath et al., 2014; Myrskylä and Fenelon, 2012; Myrskylä et al., 2017; Orlebeke et al., 1998; Tearne et al., 2015).

While the biology of ageing seems to put older parents in an unfavorable position with regard to their offspring's physical and mental, the psychosocial perspective of the effects of parental age on offspring outcomes is more nuanced. Being a child of older parents can have substantial benefits (Lawlor et al., 2011). Older parents not only are often in a better socioeconomic position than young parents (Bray et al., 2006), thereby providing a more favorable environment for children, they also have greater life experience. Furthermore, older parents display more hardiness (McMahon et al., 2007) and tend to have fewer substance use and mental health problems (Kiernan, 1997), hence score higher on parenting factors that promote health and development (Janecka et al., 2017b; Kiernan, 1997). In part, positive associations of advanced parental age could be related to selection effects. In young people, substance abuse and related externalizing problems go together with earlier sexual activity (Crockett et al., 1996), which increases the probability that intergenerational transmission of externalizing problems occurs at an early parental age (Bailey et al., 2009). Like age-related parental characteristics that may have negative effects on offspring outcomes, the influence of such selection effects can be exacerbated by assortative mating (Gratten et al., 2016).

In sum, whereas advanced parenthood, particularly advanced paternal age, has primarily been associated with physical health and neurodevelopmental outcomes, such as autism and schizophrenia, advanced parenthood, particularly advanced maternal age, rather seems to predict mental health problems with a stronger psychosocial component, such as externalizing problems. Although it seems plausible that parental age interferes with subclinical problems and traits underlying these conditions, comprehensive evidence from population-based cohorts is scarce and inconsistent, and more empirical

evidence is desirable. Moreover, prior population-based studies that used continuous measures of mental health problems usually focused on cognitive or behavioral problems (e.g., Carslake et al., 2017; Orlebeke et al., 1998) and, with a few exceptions that require replication in other cohorts (Janecka et al., 2017a; Tearne et al., 2015, 2016), rarely included internalizing problems. A final reason to extend the research conducted thus far with the present study is the wide variety of populations, designs and outcomes used, which makes it hard to distinguish between substantive variation in association patterns and sample-specific artefacts. In short, there is a need for studies that investigate both maternal and paternal age effects on continuously assessed core dimensions of offspring mental health (including internalizing problems) and that use robust analytical methods are suitable for the investigation of increased risk for both young and old parenthood.

8.0.3 The Present Study

We investigated parental age effects on offspring externalizing and internalizing problems around age 10-13 years in four Dutch population-based cohorts: Generation R (Gen-R), the Netherlands Twin Register (NTR), the Research on Adolescent Development and Relationships-Young cohort (RADAR-Y), and the Tracking Adolescents' Individual Lives Survey (TRAILS) (see Table 8.1). The Netherlands is characterized by a high maternal age at birth, and relatively few teenage pregnancies. In 1950, 1.6% of the children were born to mothers younger than 20 years of age, with a comparable percentage (1.7%) in 1990. In 2016 this number had decreased to 0.6%. In contrast, the percentages of women who gave birth at an age above 40 years were 8.5% in 1950, 1.5% in 1990, and 4.3% in 2016 (Centraal Bureau voor de Statistiek, 2018).

Table 8.1: General Cohort Information

Full cohort name	Short name	Website	Birth years	References (DOI)
Generation R	Gen-R	`generationr.nl`	2002-2006	10.1007/s10654-016-0224-9
Netherlands Twin Register	NTR	`tweelingenregister.org`	1986-2017	10.1017/thg.2012.118
Research on Adolescent Development And Relationships – Young Cohort	RADAR-Y	`www.uu.nl/onderzoek/radar`	1990-1995	10.1111/cdev.12547
TRacking Adolescents' Individual Lives Survey	TRAILS	`trails.nl`	1989-1991	10.1093/ije/dyu225

As the perception of childhood problems may differ for different informants (Rescorla et al., 2013; Hudziak et al., 2003), we aimed to obtain a comprehensive set of outcome measures of internalizing and externalizing problems through a multiple informant design. The four cohorts provided reports from mothers, fathers, the children themselves, and the children's teachers. The addition of reports from teachers is particularly valuable, because their reports are unlikely to be affected by parental age-related report biases. We tested both linear and nonlinear effects, to be better able to distinguish effects of older parenthood versus younger parenthood. We tested effects with and without adjusting for child gender and socio-economic status. Socio-economic

status was included as a covariate to get an impression of the relative importance of socio-economic factors in explaining parental age effects.

Bayesian evidence synthesis was used to summarize the results over the cohorts. The current era is one of increased awareness of the need for replication research before making scientific claims (see, for example Open Science Collaboration, 2015). Therefore, in this study, the datasets of the four cohort studies were used to evaluate the same set of hypotheses with respect to the relation between parental age and offspring mental health problems. This approach is called Bayesian evidence synthesis (Kuiper et al., 2012).

8.1 Method

8.1.1 Participants

The participants in this study came from the Gen-R, NTR, RADAR-Y, and TRAILS population cohort studies. Table 8.2 gives the total sample size and information on parental age for each cohort. The total number of children in each cohort was 4,769 for Gen-R, 25,396 for NTR, 497 for RADAR-Y, and 2,230 for TRAILS.

Table 8.2: Cohort Descriptive Statistics of Total Sample Size and Parental Age in Current Study

Cohort	N	Maternal age at birth child		Paternal age at birth child	
		Range	M (SD)	Range	M (SD)
Gen-R	4,769	16.56 - 46.85	31.68 (4.79)	17.61 - 68.67	34.24 (5.58)
NTR	25,396	17.36 - 47.09	31.35 (3.95)	18.75 - 63.61	33.76 (4.71)
RADAR-Y	497	17.80 - 48.61	31.38 (4.43)	20.34 - 52.52	33.70 (5.10)
TRAILS	2,230	16.34 - 44.88	29.32 (4.58)	18.28 - 52.09	31.99 (4.71)

Gen-R mothers were recruited in the city of Rotterdam during pregnancy. Their partners, and later their children, were also invited to participate. For Gen-R, participants from the child age-10 study wave (born between 2002 and 2006) were included if they had complete information on maternal age and a child behavioral problems sum score by at least one informant. When multiple children from one family were present, one sibling was randomly removed (N = 397) to create a sample of unrelated individuals. Mean child age for mother report was: 9.72 (SD = 0.32), father report: 9.77 (SD = 0.32), and child self-report: 9.83 (SD = 0.36). 71.2% of the Gen-R sample is Dutch or European. Other ethnic groups are Suriname (6.4%), Turkish (5.3%), and Moroccan (4.2%). Mother's educational level is low (i.e., no education or primary education) for 9%, intermediate (i.e., secondary school, lower vocational training) for 42%, and high (i.e., higher vocational training, university) for 49%. Based on mother reports, 84.5% of the children had non-clinical scores for internalizing problems, 7.1%

scored in the borderline category, and 8.4% scored in the clinical category. With respect to externalizing problems, 92.0% scored in the non-clinical category, 3.6% in the borderline category, and 4.5% in the clinical category.

The NTR study recruits new-born twins from all regions in the Netherlands. Here we included the data on 10-year-olds who were born between 1986 and 2008. Children were not included if they had a severe handicap which interfered with daily functioning. Mean child age for mother report was: 9.95 (SD = 0.51), father report: 9.94 (SD = 0.50) and teacher report: 9.80 (SD = 0.58). The children in NTR were mostly born in the Netherlands (99.5%). The remaining 0.5% consisted mainly of other West European nationalities (0.4%). Parents in the NTR were mostly born in the Netherlands (95.7% of fathers and 96.7% of mothers). The NTR genotype database indicates that 2.2% of participants born in the Netherlands have non-Dutch ancestry. 3.1% of mothers had a low skill occupation (primary education), 11.4% had an occupation that required lower secondary education, 40.3% had an upper secondary educational level, 30.6% had a higher vocational occupation level, and 14.6% worked at the highest (i.e. scientific) level. According to mother reports for internalizing problems, 86.1% of children had a non-clinical score, 5.9% had a borderline score, and 8.0% scored in the clinical range. For externalizing problems, 85.7% scored in the non-clinical range, 6.5% scored in the borderline range, and 7.8% in the clinical range.

The RADAR-Y sample was recruited in the province of Utrecht and four large cities in the mid–west of the Netherlands. Because the RADAR-Y study had a focus on delinquency development, children with borderline externalizing behavior problems at age 12 were oversampled. All participants from the first wave of data collection, born between 1990 and 1995, were selected. The mean age of the children at this wave was 13.03 years (SD = 0.46). The sample consisted mainly of native Dutch (87.9%) children. Remaining participants belonged to the following ethnic groups: Surinam (2.4%), Indonesian/ Moluccan (2.4%), Antillean (1.8%), Turkish (0.4%), and other (4.8%). Mother's educational level is low (i.e., no education or primary education) for 3.2%, intermediate (i.e., secondary school, lower vocational training) for 56.7%, and high (i.e., higher vocational training, university) for 40.1%. According to the children's reports for externalizing problems, 81.6% of the participants had a non-clinical score, 7.2% had a borderline score, and 11.2% scored in the clinical range. Using the cutoff scores for the depression scale as described by Reynolds (2000), 4.0% of the children scored in the subclinical or clinical range of depressive symptoms. Using the cutoff scores for the anxiety scale of Birmaher et al. (1997), 5.3% of the children scored in the subclinical or clinical range for anxiety symptoms.

The TRAILS sample was recruited in the Northern regions of the Netherlands. All participants from the first wave of data collection (born between 1990 and 1991) were selected. The mean age of the children at the first wave was 11.09 (SD = 0.56). The large majority of participants were Dutch (86.5%), with other participants being Surinam (2.1%), Indonesian (1.7%), Antillean (1.7%), Moroccan (0.7%), Turkish (0.5%), and other (6.9%). Mother's educational level is low (i.e., no education or primary education) for 6.9%, intermediate (i.e., secondary school, lower vocational training) for 66.3%, and high (i.e., higher vocational training, university) for 26.8%.

Based on mother-reported sum-scores for the internalizing and externalizing scales, TRAILS participants were categorized in a non-clinical, borderline, or clinical category. For internalizing problems, 67.3% of the participants had a non-clinical score, 13.9% had a borderline score, and 18.8% had a clinical score. For externalizing problems, 74.5% had a non-clinical score, 10.2% a borderline score, and 15.4% had a score in the clinical range.

To summarize, the cohorts represented the entire Dutch geographic region across all strata from society. They had a similar distribution of SES. The percentage of participants with parents born in the Netherlands was relatively high in NTR (>95%), around 87% in Radar-Y and TRAILS, and relatively low in Gen-R (<72%). The percentage of non-clinical behavioral problem scores was lowest in TRAILS.

All studies were approved by central or institutional ethical review boards. The participants were treated in compliance with the Declaration of Helsinki, and data collection was carried out with their adequate understanding and parental consent. All measures in RADAR-Y were self-reports. In the other cohorts, children were rated by any combination of: their parents, themselves, or their teachers. Table 8.3 shows the total number of children in each cohort, and the number of participants with an externalizing and internalizing behavior problem score, as a function of informant (father, mother, teacher and self).

8.1.2 Measures

Predictors

Maternal and Paternal Age at Birth. The age of the biological parents at birth of the child was measured in years up to two decimals for each cohort.

Outcomes

Externalizing and Internalizing Problems. In most cohorts, internalizing and externalizing problems were assessed by the parent-rated Child Behavior Checklist (CBCL; Achenbach and Edelbrock, 1991; Achenbach and Rescorla, 2001), the Youth Self-Report (YSR; Achenbach and Edelbrock, 1991), and the Teacher Report Form (TRF; Achenbach and Rescorla, 2001). These questionnaires contain a list of around 120 behavioral and emotional problems, which can be rated as 0 = *not true*, 1 = *somewhat or sometimes true*, or 2 = *very or often true in the past 6 months*. The broadband scale Internalizing problems includes the syndromes anxious/depressed behavior, withdrawn/depressed behavior, and somatic complaints; the broadband scale Externalizing problems involves aggressive and rule-breaking behavior. In TRAILS, the Teacher Checklist of Psychopathology (TCP) was developed to be completed by teachers. The TCP contains descriptions of problem behaviors corresponding to the syndromes of the TRF. Teachers rated the TCP on a 5-point scale (De Winter et al., 2005). In Gen-R, the YSR was replaced by the Brief Problem Monitor (BPM), containing six items for internalizing and seven items for externalizing behavior problems from the

YSR. All items were scored on a 3-point scale. In RADAR-Y, internalizing behavior problems were assessed by a combined score of the Reynolds Adolescent Depression Scale-2nd edition (RADS-2; Reynolds, 2000) and the Screen for Child Anxiety Related Emotional Disorders (SCARED; Birmaher et al., 1997) questionnaires. The RADS-2 contained 23 items (the subscale anhedonia was deleted) and the SCARED contained 38 items, which were rated on a 4-point scale (1 = *almost never*, 2 = *hardly ever*, 3 = *sometimes*, 4 = *most of the time*) and 3-point scale (1 = *almost never*, 2 = *sometimes*, 3 = *often*), respectively. Table 8.3 gives an overview of the rating instruments, the informants for each of the cohorts and the number of children in each cohort for each informant/instrument combination. A sum score was calculated per informant/instrument for the relevant items for externalizing and internalizing problems respectively. Table 8.4 shows the mean scores for externalizing and internalizing problems per cohort. The scores for girls and boys are given in Tables S1 and S2 of the supplementary materials, respectively.

Table 8.3: Total Sample Size and Sample Sizes per Rater per Cohort

(Total Sample Size)		Gen-R (N= 4,769)		NTR (N=25,396)		RADAR-Y (N=497)		TRAILS (N=2,230)	
Variable	Rater								
Externalizing behavior problems	Child	BPM[a]	4,010	-	-	YSR[b]	491	YSR[b]	2,188
	Mother	CBCL[c]	4,549	CBCL[c]	21,921	-	-	CBCL[c]	1,965
	Father	CBCL[c]	3,259	CBCL[c]	14,715	-	-	-	-
	Teacher	-	-	TRF[d]	12,573	-	-	TCP[e]	1,925
Internalizing behavior problems	Child	BPM[a]	4,018	-	-	RADS-2[f] + SCARED[g]	266	YSR[b]	2,171
	Mother	CBCL[c]	4,550	CBCL[c]	21,731	-	-	CBCL[c]	1,955
	Father	CBCL[c]	3,259	CBCL[c]	14,626	-	-	-	-
	Teacher	-	-	TRF[d]	12,389	-	-	TCP[e]	1,924

[a] Brief Problem Monitor.
[b] Youth Self Report.
[c] Child Behavior Checklist
[d] Teacher Report Form
[e] Teacher Checklist of Psychopathology
[f] Reynolds Adolescent Depression Scale - 2nd edition. Excluding anhedonia scale. Standardized before averaged with SCARED
[g] Screen for Child Anxiety Related Disorders. Standardized before averaged with RADS-2.

Covariates

Socio-Economic Status (SES) and child gender. In Gen-R, SES was defined as a continuous variable (principal component) based on parental education and household income. In NTR, SES was a 5-level ordinal variable based on occupational level. In

Table 8.4: Mean and SD for Externalizing and Internalizing Problems

Rater	Cohort	Externalizing	Internalizing	*N*-Ext/*N*-Int
Child	Gen-R	1.94 (1.92)	2.15 (2.09)	4,010/4,018
	RADAR-Y	10.61 (7.15)	-0.04 (0.86)	491/266
	TRAILS	8.68 (6.25)	11.28 (7.41)	2,188/2,171
Mother	Gen-R	3.92 (4.91)	4.86 (5.05)	4,549/4,550
	NTR	5.61 (6.12)	4.68 (5.07)	11,086/10,986
	TRAILS	8.40 (7.03)	7.85 (6.20)	1,965/1,955
Father	Gen-R	3.99 (4.91)	4.58 (4.72)	3,259/3,259
	NTR	4.66 (5.41)	3.56 (4.24)	7,420/7,374
Teacher	NTR	3.28 (5.88)	4.41 (4.96)	6,536/6,446
	TRAILS	0.44 (0.77)	0.99 (1.12)	1,925/1,924

Note. For instruments, see Table 8.3.

TRAILS, SES was a 3-level ordinal variable based on parental education, parental occupational status and household income. In RADAR-Y SES was a dichotomous variable based on parents' occupational level. Child gender was coded as male = 0 and female = 1.

Missing Data and Data Imputation

Missing Data.

For externalizing problem behavior, 15.9% of the child self-reports were missing for Gen-R, while for RADAR-Y and TRAILS these percentages were 1.2% and 1.9%, respectively. For mother reported data, 4.6% were missing for Gen-R, 13.7% for NTR and 11.9% for TRAILS. For father reported data, 31.7% were missing for Gen-R and 42.1% for NTR. For teacher reported data, 50.5% were missing for NTR and 13.7% for TRAILS. For internalizing problem behavior, the percentages were similar, except for child-reported data in RADAR-Y, where 46.4% was missing. For the predictor variables, age mother and age father, 0.3% and 1.3%, were missing for NTR, 0.0% and 14.4% for Gen-R, 0.4% and 9.7% for RADAR-Y, and 5.1% and 25.0% for TRAILS, respectively. For SES, the percentage of missing values was always below 3.0%, except for Gen-R where 22.3% was missing. For child gender, all cohorts had complete information.

Please note that the higher percentage for missing teacher- and father-reported data of NTR is due to the fact that NTR did not collect teacher-reported data at the initiation of the study and that NTR had not collected father-reported data in multiple birth years due to financial constraints. The higher percentage of missing self-reported data of internalizing problem behavior for RADAR-Y is caused by the fact that not all subscales on which the internalizing problem behavior score was based were collected from all participants.

Data Imputation.

Missing data was handled by means of multiple imputation (Schafer and Graham, 2002; Van Buuren, 2012). When multiple imputation is used, the missing values are repeatedly (in this study 100 times) imputed, that is, replaced by values that are plausible given the child's scores that are not missing, resulting in 100, so-called, completed data sets. Subsequently, each completed data set is analyzed (for example, using a multiple regression) and the 100 analyses are summarized such that the fact that "artificial data" are created by imputation is properly accounted for. Multiple imputation proceeds along three steps:

1. Determine which variables are to be used for imputation. These variables have to be chosen such that conditional on these variables the missing data are believed to be missing at random (MAR; Van Buuren, 2012), that is, whether or not a score is missing does not depend on the missing value (Schafer and Graham, 2002). Unless missingness is planned, the variables causing the missingness are unknown to the researcher. What is often done in practice is that variables are chosen that are expected to be good predictors of the variables containing missing values. One can argue with respect to which and how many variables to use, but there is no way to test whether MAR is achieved, and MAR is an assumption. The imputation model included the outcome variables externalizing and internalizing behavioral problems per informant, total behavioral problems, SES, child gender, age of the child, age of the father and age of the mother. In some cohorts, other variables were present that could also contribute to the imputation. Specifically, parent psychopathology (in Gen-R) and total number of siblings (in NTR) contributed to the imputation model. Variables functioned only as predictors when a correlation of at least .10 with the imputed variable was present. Since the NTR dataset contained twins, the imputation process differed from that of the other cohorts. The imputation for NTR was done for each family instead of each participant, so that the same value for SES, age father and age mother was obtained for both twins. The imputation of missing data was done for informants available in each cohort. So, for example, when a cohort had no teacher-reported data, teacher data were not imputed.
2. Generate imputed data matrices. The R package MICE (Multiple Imputation by Chained Equations; Van Buuren, 2012) was used to create 100 imputed data matrices. MICE uses an iterative procedure in which sequentially each variable is imputed conditional on the real and imputed values of the other variables. Continuous variables were imputed by predictive mean matching. Categorical variables were imputed using logistic regression (see Van Buuren, 2012). Success of the imputation was evaluated by checking the events logged by the software, and by checking convergence plots for a lack of trends and proper mixing of the imputation chains.
3. Analyze each imputed data set as desired and pool the results. In the current study each of the 100 imputed data sets was analyzed using multiple regression or cluster linear regression. The results, for each regression coefficient, were 100 estimates and 100 standard errors of the estimate. As may be clear, each of the standard

errors was too small because they are partly based on artificial imputed data. This was accounted for by properly pooling the results using Rubin's rules (Van Buuren, 2012, see). The variance over the 100 estimates reflects the uncertainty in the estimate due to missing values (in each of the 100 completed data sets different values are imputed). In Rubin's rules the variance of the 100 estimates is used to increase the standard errors such that they properly account for the fact that part of the data is imputed. Gen-R, TRAILS and RADAR-Y used the 'pool' function of MICE in R for summarizing the effects of the 100 separate imputed datasets, whereas NTR used the pooling option of Mplus version 8.0 (Muthén and Muthén, 2017) instead of R, to appropriately take into account the family clustering of the twins in the same analysis. Both pooling methods are based on the principles as explained here. The pooled estimates and standard errors were the main outcomes of the analyses after imputation.

8.2 Analytical Strategy: Bayesian Evidence Synthesis

The process of Bayesian evidence synthesis consists of four steps: (1) creating exploratory and confirmatory data sets; (2) generating competing hypotheses using exploratory analysis; (3) quantifying the support for each of the competing hypotheses using Bayesian hypothesis evaluation; and (4) Bayesian evidence synthesis, that is, summarizing the support resulting from each study into the overall support for the competing hypotheses in the data from the four cohort studies.

8.2.1 Exploratoration and Confirmation

8th

As was elaborated in the introduction, diverse results regarding the relation between parental age and child problem behavior have been found in the literature, with increased parental age both positively and negatively related to child problem behavior. In the same vein, there may be a quadratic effect and if there is, increased child problem behavior may be present at high and low parental age. Since research is indecisive, especially for the non-clinical studies reviewed in this paper, the data resulting from each of the cohorts were split randomly into two parts containing the same number of children: an exploratory part, which was used to generate a set of competing hypotheses; and a confirmatory part, which was used to quantify the support in the data for each of the hypotheses considered. Since the NTR dataset consisted of twins, the cross-validation datasets were split based on family ID for this cohort, to ensure independent datasets. Multiple imputation was applied separately to the exploratory and confirmatory part of the data. Having an exploratory and confirmatory dataset avoids the so-called "double dipping", that is, using *the same* data to generate and evaluate hypotheses. Here a hypothesis survived if it: 1) emerged from the exploratory analyses and 2) was supported by the confirmatory analyses. The process of generating hypotheses is explained below.

8.2.2 Generating Hypotheses using Exploratory Analyses

The exploratory half of the data resulting from each of the four cohorts was used to generate hypotheses with respect to the relation between child problem behavior and parental age. First, for each cohort separately, linear regression analyses were conducted to relate internalizing and externalizing problem behavior as evaluated by child, mother, father, and teacher (See Table 8.3 for the informants that were present per cohort) on paternal and maternal age and age squared (both with and without child gender and social economic status as covariates). For Gen-R, RADAR-Y and TRAILS, the analyses were conducted in R (R Core Team, 2017). For the NTR twin-data, cluster linear regression analyses were conducted in Mplus version 8.0 (Muthén and Muthén, 2017).

All analyses were repeated with SES and child gender as covariates. This rendered, for each combination (e.g., predicting externalizing problems as rated by the mother from mother age and age squared) an estimate of both the linear and quadratic effect for each of the cohorts that included the informant of interest. These estimates and the corresponding p-values provided information with respect to whether the linear and non-linear effects were expected to be negative, zero, or positive. To interpret the strength of relations, the variables in the exploratory analyses were all standardized. The results of the regression analyses were translated into so-called informative hypotheses (Hoijtink, 2012), that is, hypotheses that represent expectations with respect to the state of affairs in the populations from which the data of the four cohorts were sampled. An example of such an informative hypothesis is: H_1: $\beta < 0$. That is, the regression coefficient is negative. Informative hypotheses go beyond the traditional null hypothesis (here H_0: $\beta = 0$) by stating explicitly which relations between variables are expected. Often the null is added to the set of hypotheses under consideration to protect against unjustified claims that the effect specified by an informative hypothesis exists. Another hypothesis that can be added besides the informative hypotheses is the alternative hypothesis H_a: β. That is, there are no restrictions on the regression coefficient. The alternative hypothesis is used to protect against choosing the best of a set of inadequate informative hypotheses. For example, H_0: $\beta = 0$, and H_1: $\beta < 0$ constitute the set of hypotheses supported by the exploratory parts of the data, but both are inadequate in the confirmatory data. Instead, another unspecified hypothesis $\beta > 0$ describes the confirmatory data best. In this case the Bayesian approach (specified below) will prefer the alternative hypothesis, H_a: β, over the informative hypotheses H_0 and H_1. By using informative hypotheses, the exact same hypotheses could be evaluated in all cohorts, even when cohorts used different measurement instruments for the same concepts. Not requiring the exact same measurement instruments is an important benefit of Bayesian evidence synthesis over classical meta-analyses.

8.2.3 Confirmatory Bayesian Hypothesis Evaluation

Once a set of competing informative hypotheses had been formulated (including the traditional null and alternative hypotheses), the empirical support for each pair of

hypotheses was quantified using the Bayes factor (BF; Kass and Raftery, 1995). The BF is the ratio of the marginal likelihood of two competing hypotheses. Loosely speaking, the marginal likelihood of a hypothesis is the probability of that hypothesis given the data. Consequently, a BF comparing H_1 with H_a of, for example, 5 indicates that the support in the data for H_1 is five times larger than for H_a. The BF as the ratio of two marginal likelihoods implies that the fit (how well does a hypothesis describe the data set at hand) and the specificity (how specific is a hypothesis) of the hypotheses involved are accounted for (Gu et al., 2018). To give an example, if $\beta = -2$, H_1: $\beta < 0$, and H_a: β, both have an excellent fit, but H_1: $\beta < 0$ is more specific than H_a: β (anything goes), and as a result, the BF will prefer H_1 over H_a. Note that the size of the BF is related to sample size. If the precision of the evidence in the data for a hypothesis increases as a result of a larger sample, the BF for that hypothesis will increase as well. The BF implemented in the R package Bain (Gu et al., 2018) was used to evaluate informative hypotheses in the context of (cluster) multiple linear regression models.

Assuming that a priori each hypothesis is equally likely to be true, the BFs were transformed in so-called posterior model probabilities (PMPs), that is, the support in the data for the hypothesis at hand given the set of hypotheses under evaluation. PMPs have values between 0 and 1 and sum to 1 for the hypotheses in the set under consideration. For example, if PMP $H_0 = .05$, PMP $H_1 = .85$, and PMP $H_a = .10$, then it is clear that H_1 receives the most support from the data, because it has by far the largest PMP. Thus, the result of the confirmatory Bayesian hypotheses evaluation were PMPs for each hypothesis and for each informant by each of the cohorts that had ratings by this informant. The next step was to apply Bayesian evidence synthesis.

8.2.4 Bayesian Evidence Synthesis

Bayesian evidence synthesis was used to summarize the support for the hypotheses of interest over the four cohort studies. Bayesian evidence synthesis (Kuiper et al., 2012) can be illustrated using the set of hypotheses: H_0: $\beta = 0$, H_1: $\beta < 0$, and H_a: β as introduced above. In the context of this paper, these hypotheses are incompletely specified. The complete specification would be H_0: $\beta = 0$ for NTR, H_1: $\beta < 0$for NTR and H_a: β for NTR, and analogously for the other three cohort studies. This specification highlights that the support for the hypotheses depends on the cohort study at hand. Bayesian evidence synthesis can then be used to determine support for a set of hypotheses:

H_0:H_0 for NTR & H_0 for TRAILS & H_0 for Gen-R & H_0 for Radar-Y
H_1:H_1 for NTR & H_1 for TRAILS & H_1 for Gen-R & H_1 for Radar-Y
H_a:H_a for NTR & H_a for TRAILS & H_a for Gen-R & H_a for Radar-Y

that is, the regression coefficient is zero *in the populations corresponding to each of the four cohort studies*, the regression coefficient is smaller than zero *in the populations corresponding to each of the four cohort studies*, and there is no prediction with respect to the regression coefficient *in the populations corresponding to each of the*

four cohort studies. If for a specific set of hypotheses only two or three cohorts contain the necessary variables, the hypotheses can be adjusted accordingly. Like for each individual study, the support for these composite hypotheses was quantified using posterior model probabilities (PMPs).

If a hypothesis emerges from the exploratory analyses of the data corresponding to the cohort studies and is supported by the confirmatory analyses of the data corresponding to the cohort studies, then there is evidence that this hypothesis provides an adequate description of the relation between child problem behavior and parental age, that is, in general, independent of the specific cohort studies used to evaluate this hypothesis. With the methodological approach elaborated in this section and applied in the remainder of this paper, the increased awareness of the need for replication studies before making scientific claims is explicitly addressed.

8.3 Results

8.3.1 Exploratory Analyses

The results of the exploratory analyses (see Supplementary Materials) generally showed a negative relation between mean-centered parental age and externalizing problems accompanied by a positive quadratic coefficient, implying that the negative relation with age at the mean declined across age (see Table S3 and Figure S1). This model explained about 1.9% of the total variance in externalizing problems with maternal age and 1.2% with paternal age. For internalizing problems, the relation with parental age was less apparent: about 0.5% of the total variance was explained by maternal age, and about 0.2% was explained by paternal age. In analyses including the covariates SES and gender, the relation with age diminished, but remained significant (Tables S4, and S5 of the Supplementary Materials). Higher SES was related to fewer externalizing problems, and boys showed more externalizing problems than girls. In general, no relation between parental age and internalizing problems was observed (see Tables S6, S7, and S8, and Figure S1 of the Supplemental Materials).

Our interpretation of the exploratory results led to the following set of competing informative hypotheses with respect to the relation between parental age (mean-centered), as indicated by a linear (i.e., β_1) and quadratic (i.e., β_2) coefficient, and child problem behavior:

H_1:$\beta_1 = 0$, $\beta_2 = 0$. Age does not have a linear or quadratic relation.
H_2:$\beta_1 < 0$, $\beta_2 = 0$. Age has a negative linear relation, there is no quadratic relation.
H_3:$\beta_1 < 0$, $\beta_2 > 0$. Age has a negative linear relation, and a positive quadratic relation.
H_4:β_1, β_2. The coefficients can have any value.

Based on the exploratory results, we expected most evidence for H_2 or H_3 in analyses with parental age predicting externalizing problems, and most evidence for H_1 in analyses with parental age predicting internalizing problems.

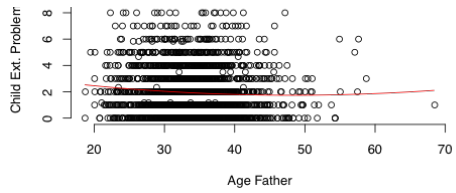
Since the exploratory results did not show a positive linear or a negative quadratic relation between age and behavioral problems, the hypotheses do not include these features. However, we remained open to other options by including the alternative hypothesis H_a that imposes no constraints on the parameters, and accordingly claims that anything can be true. H_a receives the most support if none of the specified informative hypotheses provides an adequate description of the confirmatory part of the data from each of the four cohorts. In this manner, we avoided that the best hypothesis out of the set of H_1, H_2, and H_3, is an implausible hypothesis.

8.3.2 Confirmatory Analyses

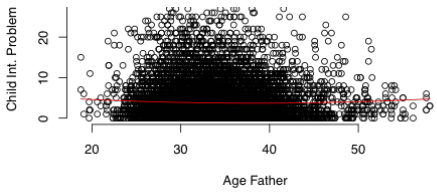
Tables S9 to S14 contain the confirmatory unstandardized regression coefficients. These are the results per cohort that generated the relative support for the competing informative hypotheses as will be presented in the next paragraph. We will discuss the underlying results briefly. Similarly to the exploratory data, the results showed negative relations across cohorts between parental age and externalizing problems. However, in the confirmatory data, the quadratic coefficients from the cohorts were less often significantly different from zero than in the exploratory data. The model with a linear and quadratic coefficient for parental age explained on average about 1.1% of the total variance in externalizing problems with maternal age and 0.9% with paternal age as a predictor. With respect to internalizing behavior problems, the model with maternal age explained on average about 0.4% of the total variance, and paternal age explained on average about 0.3%. Figure 8.1 visualizes the relation between age and behavioral problems using the first imputation of the confirmatory part of Gen-R and NTR respectively. The figure presents a plot of data for internalizing and externalizing problems. As a result of centering, the linear effect that we investigated is the effect at the mean age around 29-32 years for mothers and 32-34 years for fathers (see Table 8.2 for mean parental age per cohort). The results presented in the figures were representative for all other analyses and cohorts.

8.3.3 Parental Age and Externalizing Behavior Problems

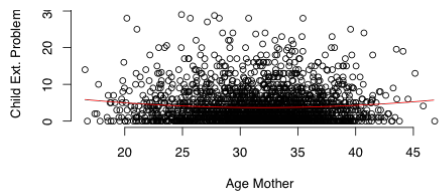
The posterior model probabilities (PMPs) concerning the relation between parental age and externalizing problems are presented in Table 8.5. The table only shows PMP scores for those cohorts that included the associated informants (see Table 8.3 for an overview of informants per cohort). As shown in Table 8.5, for parent-reported externalizing behavior problems, Gen-R yielded most evidence for H_1 (i.e., no relation with parental age); NTR mostly supported H_2, (i.e., the relation with parental age is linear and negative) as did TRAILS, but for mother-reported externalizing behavior problems predicted by paternal age, NTR yielded most support for H_3 (i.e., the relation with parental age follows a negative linear trend including a positive quadratic factor). The combined results for mother-reported externalizing behavior problems predicted by father age showed substantial support (PMP = .53 and .45 respectively) for H_2 and H_3. For father reported externalizing behavior problems predicted by father age



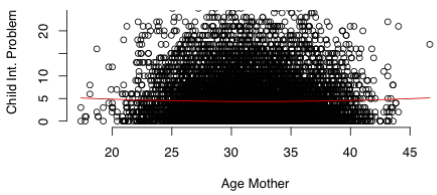
(a) Gen-R child-reported externalizing problems in relation to paternal age



(b) NTR father-reported internalizing problems in relation to paternal age



(c) Gen-R mother-reported externalizing problems in relation to maternal age



(d) NTR teacher-reported internalizing problems in relation to maternal age

Fig. 8.1: Confirmatory results for parental age in relation to problem behavior as represented in Gen-R and NTR.

and for parent-reported externalizing behavior problems predicted by mother age, the combined results provided most support for H_2: the relation with parental age is linear and negative, in other words, higher parental age is associated with less externalizing behavioral problems. For teacher-reported externalizing behavior problems predicted by paternal age, TRAILS and NTR combined yielded most evidence for H_1 (i.e., no relation with parental age) closely followed by H_2. When maternal age was included, most support was found for H_2: the relation with parental age is linear and negative. For child-reported externalizing behavior problems, the results were mixed across cohorts (Gen-R preferred H_2 or H_3, RADAR-Y H_3 or H_1, and TRAILS H_1). After combining the results from the three cohorts, however, most support was obtained for H_1, that is, no relation with parental age.

Table 8.6 shows the results after inclusion of the covariates as predictors of externalizing problems. After adjusting for SES and gender, all cohorts yielded substantial evidence for H_1 with respect to child- and teacher-reported externalizing problem behavior. This meant a shift especially for the child-reported problem behavior by Gen-R, and the teacher-reported problem behaviors by both NTR and TRAILS. For parent-reported problem behavior, some cohorts provided most support for H_1 (Gen-R for all parent-reports, and TRAILS for paternal age predicting mother-reported problem behavior), others for H_2 (TRAILS and NTR), and NTR for H_3 in mother-reported problem scores related to paternal age. By including covariates in the model, Gen-R

and TRAILS mainly handed in support on H_2 while in NTR the support for H_2 increased at the expense of support for H_3. When combining evidence for the parent reports, most support was still found for H_2, that is, there is a linear inverse relation between parental age and externalizing problem behavior.

8.3.4 Parental Age and Internalizing Behavior Problems

With regard to internalizing problems (the results are presented in Table 8.7), the cohorts generally found most evidence for H_1 for multiple informants, except for mother-reported internalizing problems reported by maternal age in NTR. All combinations of studies rendered most support for H_1, which means that the hypothesis that there is no relation between parental age and internalizing problems was best supported by the set of studies.

After including the covariates SES and gender (Table 8.8), all results still suggested the most support for H_1 for the impact of parental age on internalizing problem behavior, irrespective of the cohort and informant. Consequently, combining the results from the various cohorts provided overwhelming support for H_1, that is, there is no evidence for a relation between parental age and child internalizing problem behavior.

Table 8.5: Posterior Model Probabilities for Parental Age Predicting Externalizing Problems

Rater	Cohort	Age Father				Age Mother			
		H_1	H_2	H_3	H_a	H_1	H_2	H_3	H_a
Child	Gen-R	.23	*.56*	.16	.05	.22	.18	*.49*	.13
	RADAR-Y	.28	.02	*.49*	.22	*.43*	.07	*.38*	.12
	TRAILS	*.86*	.13	.00	.01	*.83*	.15	.02	.01
	All	**.98**	.02	.00	.00	**.93**	.02	.04	.00
Mother	Gen-R	*.90*	.07	.02	.01	*.82*	.04	.10	.05
	NTR	.00	.02	*.74*	.24	.00	*.89*	.09	.03
	TRAILS	.18	*.74*	.06	.02	.00	*.88*	.09	.03
	All	.00	**.53**	**.45**	.00	.00	**.97**	.03	.00
Father	Gen-R	*.65*	.22	.10	.03	*.60*	.19	.17	.04
	NTR	.00	*.49*	.38	.13	.00	*.93*	.05	.00
	All	.00	**.73**	.25	.02	.00	**.95**	.05	.00
Teacher	NTR	*.55*	.41	.03	.01	.29	*.60*	.09	.02
	TRAILS	*.48*	.31	.16	.05	.00	*.73*	.21	.06
	All	**.67**	.32	.01	.00	.00	**.96**	.04	.00

Note. Numbers in italic font represent the highest posterior model probability per cohort. Numbers in bold font represent the highest meta-analytic results.

Table 8.6: Posterior Model Probabilities for Parental Age Predicting Externalizing Problems after Correction for Impact Covariates

Rater	Cohort	Age Father				Age Mother			
		H_1	H_2	H_3	H_a	H_1	H_2	H_3	H_a
Child	Gen-R	*.62*	.33	.04	.01	*.83*	.10	.05	.02
	RADAR-Y	*.36*	.02	*.42*	.19	*.53*	.08	.29	.10
	TRAILS	*.88*	.11	.00	.01	*.89*	.09	.02	.01
	All	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00
Mother	Gen-R	*.96*	.03	.00	.00	*.97*	.02	.00	.01
	NTR	.00	.31	*.52*	.17	.00	*.95*	.04	.01
	TRAILS	*.67*	.31	.01	.01	.30	*.63*	.05	.02
	All	.03	**.99**	.00	.00	.00	**1.00**	.00	.00
Father	Gen-R	*.88*	.10	.02	.00	*.92*	.06	.01	.00
	NTR	.02	*.84*	.02	.00	.00	*.96*	.03	.01
	All	**.72**	.28	.00	.00	.00	**.99**	.01	.00
Teacher	NTR	*.79*	.20	.01	.00	*.68*	.28	.03	.01
	TRAILS	*.87*	.11	.02	.00	*.60*	.32	.07	.02
	All	**.97**	.03	.00	.00	**.81**	.18	.00	.00

Note. Numbers in italic font represent the highest posterior model probability per cohort. Numbers in bold font represent the highest meta-analytic results.

Table 8.7: Posterior Model Probabilities for Parental Age Predicting Internalizing Problems

Rater	Cohort	Age Father				Age Mother			
		H_1	H_2	H_3	H_a	H_1	H_2	H_3	H_a
Child	Gen-R	*.91*	.08	.01	.00	*.86*	.09	.04	.01
	RADAR-Y	*.84*	.09	.05	.03	*.81*	.16	.02	.01
	TRAILS	*.96*	.04	.00	.00	*.93*	.06	.01	.00
	All	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00
Mother	Gen-R	*.58*	.25	.14	.04	*.35*	.25	*.33*	.08
	NTR	*.69*	.26	.04	.01	.26	*.72*	.01	.01
	TRAILS	*.94*	.05	.00	.00	*.80*	.17	.02	.01
	All	**.99**	.01	.00	.00	**.71**	.29	.00	.00
Father	Gen-R	*.43*	*.42*	.11	.03	*.48*	.36	.13	.03
	NTR	*.96*	.04	.00	.00	*.95*	.05	.00	.00
	All	**.96**	.04	.00	.00	**.97**	.03	.00	.00
Teacher	NTR	*.99*	.01	.00	.00	*.99*	.01	.00	.00
	TRAILS	*.85*	.06	.07	.02	.24	.15	*.49*	.12
	All	**1.00**	.00	.00	.00	**.99**	.01	.00	.00

Note. Numbers in italic font represent the highest posterior model probability per cohort. Numbers in bold font represent the highest meta-analytic results.

Table 8.8: Posterior Model Probabilities for Parental Age Predicting Internalizing Problems after Correction for Impact Covariates

Rater	Cohort	Age Father				Age Mother			
		H_1	H_2	H_3	H_a	H_1	H_2	H_3	H_a
Child	Gen-R	*.77*	.21	.02	.01	*.82*	.09	.07	.02
	RADAR-Y	*.86*	.07	.04	.03	*.86*	.11	.02	.01
	TRAILS	*.97*	.03	.00	.00	*.95*	.04	.00	.00
	All	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00
Mother	Gen-R	*.88*	.11	.01	.00	*.93*	.05	.01	.00
	NTR	*.88*	.11	.01	.00	*.70*	.29	.00	.00
	TRAILS	*.96*	.04	.00	.00	*.91*	.08	.01	.00
	All	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00
Father	Gen-R	*.88*	.09	.02	.01	*.90*	.08	.01	.00
	NTR	*.96*	.03	.00	.00	*.96*	.04	.00	.00
	All	**1.00**	.01	.00	.00	**1.00**	.01	.00	.00
Teacher	NTR	*.99*	.01	.00	.00	*.99*	.01	.00	.00
	TRAILS	*.94*	.04	.02	.01	*.83*	.06	.08	.03
	All	**1.00**	.00	.00	.00	**1.00**	.00	.00	.00

Note. Numbers in italic font represent the highest posterior model probability per cohort. Numbers in bold font represent the highest meta-analytic results.

8.4 Discussion

8.4.1 Parental Age and Externalizing Problems

We found evidence for a negative linear relation between parental age and externalizing problems as reported by parents. That is, older parents have children with less externalizing behavior problems. There was also evidence for a negative linear relation between maternal age and externalizing problems as reported by teachers. For teachers, this finding was partly explained by SES. However, the relation between parental age and parent-reported externalizing problems persisted after adjusting for SES, so the favorable effect of parental age is not solely due to SES.

8.4.2 Parental Age and Internalizing Problems

Parental age seemed unrelated to child internalizing problem behavior, especially when accounting for SES. Tentatively, older parenthood might be associated with both high and low vulnerability to develop internalizing problems. On the one hand, older parents may have a lower probability of internalizing problems because they are less likely to have a background characterized by deprivation and social instability (Robson and Pevalin, 2007), known to be related to internalizing problems such as anxiety and depression. On the other hand, internalizing problems can increase the probability of older parenthood, by hampering engagement in and consolidation of romantic relationships (Manning et al., 2010; Sandberg-Thoma and Dush, 2014). Possibly, both processes play a role, and their joint influence results in a lack of net result.

8.4.3 Sociodemographic Factors as a Potential Explanation

The relatively consistent beneficial effect of advanced parenthood for childhood externalizing problems may seem unexpected, given mixed findings from earlier research on more common mental health problems (de Kluiver et al., 2017; McGrath et al., 2014). The beneficial effect of advanced parental age could have more than one explanation. Older and younger parents have different parenting styles. For example, there is evidence that older mothers use less frequent sanctions towards their children, are more sensitive to the child's needs and provide more structure (Trillingsgaard and Sommer, 2018). Older parents may also tend to appraise a specific problem level as less disturbing than younger parents, and older parents might be more patient and are capable of setting limits, thus feeling more equipped to handle externalizing behaviors. The positive impact of higher quality parenting by older parents is expected to be more relevant to externalizing problem behavior than to autism and schizophrenia, where a disadvantageous impact of increased parental age has been established.

Previous studies provided evidence indicating that offspring of older parents are, in several respects, more affluent than those with younger parents (e.g., Carslake et al., 2017; McGrath et al., 2014; Myrskylä and Fenelon, 2012; Orlebeke et al., 1998; Tearne et al., 2015, 2016). The finding that the negative relation of parental age and externalizing problems became weaker when SES was taken into account, indicates that the relatively high SES of older parents, or SES-related selection effects (Robson and Pevalin, 2007) at least partly explained why their children have a decreased probability of externalizing problems. Myrskylä et al. (2017) argued that there are indeed important socio-demographic pathways associated with delayed parenthood in more recent birth cohorts. Older mothers tend to have better health behaviors during pregnancy, for example with respect to smoking during pregnancy, which is an established risk factor for offspring externalizing problems (Dolan et al., 2016).

Furthermore, parents who have externalizing behavior problems themselves may be higher in risk taking and may have children at a younger age. Hence, externalizing behavior problems may be transmitted especially by younger parents and less by older parents. This idea is in line with the unclarity about a relation between ADHD and advanced paternal age (de Kluiver et al., 2017; McGrath et al., 2014).

From a biological point of view, advanced parenthood seems mostly disadvantageous, but socio-demographic factors might compensate (or even more than compensate) for the biological disadvantages related to reproductive ageing when it comes to mental health problems. Older mothers from more recent birth cohorts are more socioeconomically advantaged, and happier after childbearing. The observation that older parents have offspring with fewer externalizing problems, tended to disappear when SES was taken into account. This shows that demographic factors can indeed compensate for the biological disadvantages.

8.4.4 Earlier Versus Later Birth Cohorts

In the 1950s and 1960s the number children born to mothers over the age of 40 was larger than in 2016. For offspring born during the 1960s, Saha et al. (2009) found

a negative association between maternal age and externalizing behavior problems, but in contrast to our results, they observed a positive association between maternal age and internalizing problems, and a positive association between paternal age and externalizing behavior problems. The study differed in several important aspects from the current one. All offspring were born during the 1960s, whereas in our study, all offspring were born after 1980. The age at which fathers and mothers have children has increased in the last 20 years. In the Saha et al. study average maternal and paternal ages were 24.8 and 28.4, respectively, while in our samples average maternal- and paternal ages were around 31 and 33 years. Older mothers from earlier birth cohorts tended to have low levels of education and their offspring had many older siblings (Myrskylä et al., 2017). In later birth cohorts, older mothers had higher education than younger mothers and their offspring had fewer older siblings. Thus, the family resources are spread less thinly across siblings than in earlier times. This may be the reason that our results differ from some of the findings of Saha et al. (2009). As argued by Myrskylä et al. (2017) as well, being a parent during the 1960s differs from being a parent in the 1980s, and children born during the 1980s and later might benefit from positive changes in the macro-environment.

8.4.5 Informant Effect

We used a multi-informant design (i.e., mother, father, teacher, child) to investigate parental age effects on behavioral problems. Most questionnaires belonged to the same system (ASEBA), but they do not necessarily capture the exact same construct, as different informants observe the children in different contexts. It is well-established that correlations between different types of informants are modest at the most (Achenbach et al., 1987; Renk and Phares, 2004), and it is generally recommended to involve multiple informants to assess child and adolescent psychopathology (Jensen et al., 1999). Consistent with the notion that different informants provide partly non-overlapping information, the results in this study depended on the choice of informant, since, as opposed to parent-reported problems, child-reported externalizing problems were not predicted by parental age. Conceivably, this different outcome for child-reported problems is due to a limited ability of 10-year-old children to report reliably and validly on their externalizing behaviors. It is less likely that the associations with parent-reports are caused by reporter bias because, as teacher-reports also provided support for an association with maternal age. Thus, the choice of informant is not an arbitrary one, and may influence the associations that are found. Obviously, the parent and teacher sample sizes were also substantially larger than the sample size for child-reports. Additionally, the largest study with child reports (i.e., TRAILS) used a shortened version of the YSR, which could cause lower reliability and validity of child-reports.

8.4.6 Strengths of the Current Paper

This paper adopted an analysis strategy that used the data of multiple cohort studies to evaluate the same set of hypotheses. First, the data of each cohort study were divided

into two parts: an exploratory part and a confirmatory part. Second, the exploratory part was used to generate a set of competing informative hypotheses. Third, the confirmatory part was used to compute the support in each cohort for the hypotheses entertained and to combine studies by means of Bayesian updating to compute overall results (Kuiper et al., 2012). This analysis strategy had a number of advantages. In the exploratory analyses data snooping or even p-hacking is allowed, because this part of the data is only used to generate a set of competing informative hypotheses and not to evaluate these hypotheses. In contrast, the confirmatory part of each data set is only used to evaluate this set of informative hypotheses to the traditional null and alternative hypotheses, which should, especially in ages of replication crisis, publication bias and questionable research practices, increase the credibility of our results. The interested reader is referred to the Supplementary Materials where we highlight why exploratory analyses may lead to incorrect interpretations, even with large samples, and that cross-validation can prevent this from happening. In addition, with traditional null hypothesis significance testing, we would not have been able to quantify the support for the null hypothesis (p-values cannot be used to "accept" the null-hypothesis), which appeared an important hypothesis in our study. Bayes factors and posterior model probabilities are not used to reject or not reject the null-hypotheses, they are used to quantify the support in each of the cohorts for the hypotheses entertained. Furthermore, combining studies using Bayesian updating enabled us to quantify the relative evidence with respect to multiple hypotheses using the data from multiple cohorts. Again, in ages of replication crisis, it is valuable to base conclusions on data from multiple cohorts that can all be used to address the same research question.

8.4.7 Limitations

Although the study has a number of methodological strengths, there are also limitations. First, the study focused on children's externalizing and internalizing behavior problems and did not examine other outcomes that may be positively associated with parental age, such as physical health problems and neurodevelopmental conditions. Second, children's behavior problems were only assessed during early adolescence. Thus, the study could not investigate the possibility that the direction or magnitude of the associations may vary at different points in development. For example, previous research suggesting a negative association between parental age and individuals' well-being has focused on late adolescents and young adults (e.g., Tearne et al., 2016; Weiser et al., 2008). Third, a tiny percentage of the parents were under the age of 20 at the time of the child's birth. Although this reflects societal changes in the Netherlands, it would be important to note that some results may not replicate in other populations that have higher percentages of teenage pregnancies. This may be especially relevant when interpreting the lack of an association between parental age and children's internalizing behavior problems in this study.

8.4.8 Conclusion

The analytic strategy applied to large cohorts showed us a beneficial association between advanced parental age and externalizing problem behavior, while for internalizing problem behavior there was no beneficial association with parental age. We found no evidence for a harmful effect of advanced parenthood.

9

Discussion

Research makes the greatest progress when it makes use of the results and insights of others. Hence, the motto of Google Scholar: "Stand on the shoulders of giants". Informative prior distributions and informative hypotheses provide ways to formalize and include the findings of others in a statistical analysis.

The aim of this dissertation was to demonstrate how prior knowledge can be formalized and evaluated. In Part I, the emphasis was on the formalization of prior knowledge. Chapter 2 clarified that prior knowledge can promote convergence, coverage, correct estimates, and statistical power. Especially for smaller subgroups in multi-group models, prior knowledge can make an important contribution. The empirical basis of the simulation study in Chapter 2 makes the results directly applicable to Chapter 3. Researchers that want to explore the impact of specific prior variances in their research context, can make use of the simulation set up used in this chapter.

Chapter 3 shows that it is not easy to acquire the prior information in practice. Meta-analyses that may provide the most useful information may not be available, and reviews provide qualitative instead of quantitative information, which makes it harder to transform the information into prior distributions. Subgroups in the model for which it is difficult to obtain participants will often also be the groups for which it is hard to acquire prior information, while Chapter 2 demonstrated that prior information for these groups is most important. Under these challenging circumstances, Chapter 3 pointed out that individual studies, experts and general knowledge can proof to be useful resources for prior information.

Constructing prior distributions from the obtained prior information is not straightforward: decisions need to be made about the type of distribution, its mean, mode and variance. Visualizations of the distributions can help researchers in the decision making process. It is important that the researcher is clear about the prior specifications and the reasoning behind them. Next to the empirical application, Chapter 3 provides guidelines for researchers that want to include informative priors in their analyses, which filled an important gap in the current literature. Zweers (2018, Chapter 5) followed the guidelines in Zondervan-Zwijnenburg et al. (2017a) and could conduct successful analyses with the acquired prior information. More applications would provide

This chapter is written by Mariëlle Zondervan-Zwijnenburg

additional insight in the process of acquiring prior information and help researchers further in using prior information.

In Chapter 4, we developed and evaluated an expert elicitation procedure for correlations. The elicitation procedure appears promising, and a beneficial property of the bins and chips method is that it results in complete distributions that can be used as priors in Bayesian analyses. Note, however, that expert elicitation procedures always have to be customized to the expert (group) at hand. The procedure should be introduced properly and should match the statistical and content knowledge of the specific experts. Veen et al. (2017) continued the work of Zondervan-Zwijnenburg et al. (2017b) by building an online trial roulette method with immediate feedback: it shows how the expert's decisions affect the shape of the prior distribution. This elicitation method may further improve the validity and reliability of experts' elicited knowledge, and can easily be incorporated in an elicitation procedure. The preferred elicitation mode (e.g., online, face-to-face, in a meeting with computers) again depends on the expert group at hand. As was implied by the retest in Chapter 4, some experts may focus better on the elicitation during a meeting than while at work or at home.

All in all, Part I of this dissertation showed that prior information can be very useful, but it comes with a price. It can take substantial effort to obtain prior information. Furthermore, researchers need to make many decisions in the process, for example, about the form of the distribution, or about the customization of the expert elicitation procedure. Bayesian statistics are sometimes criticized because informative priors would be subjective. The fact that prior distributions contain information independent of the data, however, does not make them unscientifically subjective. The fact that decisions are involved in the process of constructing the priors may endanger objectivity, but not more than objectivity is endangered by researcher degrees of freedom in standard data collection and analysis: In every study, researchers need to make decisions on the operationalizations of constructs, measurement instruments, the statistical model etcetera (see also Chapter 7). Decision making is part of scientific research, but to protect integrity, decisions should be independent of outcomes and need to be clarified. This transparency is encouraged throughout the current dissertation.

Part II is about testing replication. In Chapter 5 and Chapter 6, the prior predictive p-value is introduced to test whether a new study fails to replicate relevant findings of an original study. The prior predictive p-value takes into account that parameter estimates and samples based on these estimates vary naturally. In contrast to what some replication assessment methods would conclude, not finding the same results is not necessarily evidence for non-replication. Only if the new results are extreme in comparison to datasets predicted given the original results, the prior predictive p-value concludes that studies do not replicate. Furthermore, the prior predictive p-value is applicable to a wide range of statistical models. Testing the replication of claims of an original study beyond an effect size (e.g., Cohen's d or Pearson's r) was not possible with any of the previously proposed replication testing methods. The extension of the test beyond the ANOVA model in Chapter 6 makes an important contribution to the replication literature in this respect. An R Shiny user-interface, and two R-packages were developed to make the prior predictive p-value accessible to

social and behavioral scientists. Future research may further investigate how a single prior predictive p-value can be obtained in case of missing data in the new study. Furthermore, statistical power is important for the prior predictive p-value. For the ANOVA model, we developed a method to compute statistical power and calculate the required sample size for a sufficiently powered replication study. Future research may delve deeper into defining the alternative population for which replication should be rejected for models beyond the ANOVA. If the alternative population is defined, a method to calculate power and required sample sizes can be developed accordingly.

Although the prior predictive p-value is useful in the replication context, Chapter 7 shows that there are also other replication questions that can be answered with different methods. For example, researchers may want to investigate the similarity of the original and new study, or may want to focus on the population effect. For the first question Bayes factors can be useful, while the population effect may be best tackled with a meta-analytic approach that takes publication bias into account. Chapter 7 discussed four replication research questions and associated methods with special attention to small sample research. For many of the methods related to alternative replication research questions, an extension to models beyond the correlation and t-test is warranted.

Finally, in Part III, we used exploratory results to compose informative hypotheses that were tested in four cohort studies. By using separate exploratory and confirmatory datasets, we reduced the risk of overfitting, that is, modeling irregularities of a sample and interpreting them as population effects. The confirmatory analyses resulted in a posterior model probability for each of the competing informative hypotheses. Afterwards, the relative probabilities were combined over the four cohorts to evaluate which hypotheses were best supported by all cohorts. The four cohorts used varying measures for internalizing and externalizing behavior problems. Hence, with a classic meta-analysis it would not have been possible to combine the different studies. By using informative hypothesis, we evaluated which hypothesis was robustly supported by all cohorts, irrespective of measurement choices and sample characteristics. Chapter 8 demonstrated that evaluating and updating informative hypotheses is a good and useful method for research synthesis and is ready to be applied. Future directions may concern prior weights that can be allocated to the involved studies or hypotheses, and the application to, for example, longitudinal structural equation models.

References

Achenbach, T. M. and Edelbrock, C. (1991). *Child behavior checklist*. University of Vermont.

Achenbach, T. M., McConaughy, S. H., and Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101(2), 213–232. doi: 10.1037/0033-2909.101.2.213.

Achenbach, T. M. and Rescorla, L. (2001). *Manual for the ASEBA school-age forms & profiles*. University of Vermont.

Achterberg, M., van Duijvenvoorde, A. C. K., Van der Meulen, M., Bakermans-Kranenburg, M. J., and Crone, E. A. (2018). Heritability of aggression following social evaluation in middle childhood: An fMRI study. *Human Brain Mapping*, 39(7), 2828–2841. doi: 10.1002/hbm.24043.

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

Albert, J. (2009). *Bayesian computation with R*. Springer Science & Business Media. doi: 10.1007/978-0-387-92298-0.

Albert, J. (2014). *LearnBayes: Functions for Learning Bayesian Inference*. R package version 2.15. Available from: http://CRAN.R-project.org/package=LearnBayes.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders (DSM)*. American Psychiatric Association, Washington, DC, 4 edition.

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

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ..., and Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27(2), 108–119. doi: 10.1002/per.1919.

Asparouhov, T. and Muthén, B. O. (2010). Bayesian analysis of latent variable models using Mplus. Available from: http://www.statmodel.com/techappen.shtml.

Bailey, J. A., Hill, K. G., Oesterle, S., and Hawkins, J. D. (2009). Parenting practices and problem behavior across three generations: Monitoring, harsh discipline, and drug use in the intergenerational transmission of externalizing behavior. *Developmental Psychology*, 45(5), 1214–1226. doi: 10.1037/a0016129.

Bakker, A., Van der Heijden, P. G., Van Son, M. J., and 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. doi: 10.1037/a0033983.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4), 1281–1312.

Barnes, G. E., Barnes, M. D., and Patton, D. (2005). Prevalence and predictors of "heavy" marijuana use in a Canadian youth sample. *Substance Use and Misuse*, 40(12), 1849–1863. doi: 10.1080/10826080500318558.

Bertolino, F. and Racugno, W. (1992). Analysis of the linear correlation coefficient using pseudo-likelihoods. *Journal of the Italian Statistical Society*, 1(1), 33–50. doi: 10.1007/bf02589048.

Best, J. R. and Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, 81(6), 1641–1660. doi: 10.1111/j.1467-8624.2010.01499.x.

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., and Neer, S. M. (1997). The screen for child anxiety related emotional disorders (scared): Scale construction and psychometric characteristics. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(4), 545–553. doi: 10.1097/00004583-199704000-00018.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods*, 14(3), 225–238. doi: 10.1037/a0016619.

Boomsma, A. and Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In Cudeck, R., Jöreskog, K. G., and Sörbom, D., editors, *Structural equation models: Present and future. A Festschrift in honor of Karl Jöreskog*, pages 139–168. Scientific Software International.

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. doi: 10.2307/2982063.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., ..., and Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, 50, 217–224. doi: 10.1016/j.jesp.2013.10.005.

Bray, I., Gunnell, D., and Smith, G. D. (2006). Advanced paternal age: How old is too old? *Journal of Epidemiology & Community Health*, 60(10), 851–853. doi: 10.1136/jech.2005.045179.

Can, S., Van de Schoot, R., and Hox, J. (2015). Collinear latent variables in multilevel confirmatory factor analysis a comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement*, 75(3), 406–427. doi: 10.1177/0013164414547959.

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

Carslake, D., Tynelius, P., van den Berg, G., Smith, G. D., and Rasmussen, F. (2017). Associations of parental age with health and social factors in adult offspring. methodological pitfalls and possibilities. *Scientific Reports*, 7(1). doi: 10.1038/srep45278.

Centraal Bureau voor de Statistiek (2018). Geboorte; kerncijfers. Available from: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37422ned/table?ts=1522410899684.

Chandler, J. (2015). Replication of Janiszewski & Uy (2008, PS, study 4b). online. Available from: osf.io/aaudl.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017). *shiny: Web Application Framework for R*. R package version 1.0.5. Available from: https://CRAN.R-project.org/package=shiny.

Clarke, T. L. (2009). *Executive Functions and Overt/Covert Patterns of Conduct Disorder Symptoms in Children With ADHD*. PhD thesis, University of Maryland.

Clemen, R. T., Fischer, G. W., and Winkler, R. L. (2000). Assessing dependence: Some experimental results. *Management Science*, 46(8), 1100–1115. doi: 10.1287/mnsc.46.8.1100.12023.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145–153. doi: 10.1037/h0045186.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edition. doi: 10.4324/9780203771587.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. doi: 10.1037/0003-066X.49.12.997.

Cooke (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. New York: Oxford University Press.

Cooke, R. M. and Goossens, L. (1999). Procedures guide for structured expert judgment. *Brussels-Luxembourg: Commission of the European Communities*.

Crockett, L. J., Bingham, C. R., Chopak, J. S., and Vicary, J. R. (1996). Timing of first sexual intercourse: The role of social control, social learning, and problem behavior. *Journal of Youth and Adolescence*, 25(1), 89–111. doi: 10.1007/bf01537382.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3(4), 286–300. doi: 10.1111/j.1745-6924.2008.00079.x.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29. doi: 10.1177/0956797613504966.

de Kluiver, H., Buizer-Voskamp, J. E., Dolan, C. V., and Boomsma, D. I. (2017). Paternal age and psychiatric disorders: A review. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 174(3), 202–213. doi: 10.1002/ajmg.b.32508.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, 18, 186–219. doi: 10.1037/a0031609.

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

Desjardins, B., Bideau, A., and Brunet, G. (1994). Age of mother at last birth in two historical populations. *Journal of Biosocial Science*, 26(04). doi: 10.1017/s0021932000021635.

Dolan, C. V., Geels, L., Vink, J. M., van Beijsterveldt, C. E. M., Neale, M. C., Bartels, M., and Boomsma, D. I. (2016). Testing causal effects of maternal smoking during pregnancy on offspring's externalizing and internalizing behavior. *Behavior Genetics*, 46(3), 378–388. doi: 10.1007/s10519-015-9738-2.

Earp, B. D. and Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6, 621. doi: 10.3389/fpsyg.2015.00621.

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

Etz, A. and Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLOS ONE*, 11(2), e0149794. doi: 10.1371/journal.pone.0149794.

Fischer, P., Greitemeyer, T., and Frey, D. (2008). Self-regulation and selective exposure: The impact of depleted self-regulation resources on confirmatory information processing. *Journal of Personality and Social Psychology*, 94(3), 382. doi: 10.1037/0022-3514.94.3.382.

Fontes, M., Bolla, K., Cunha, P., Almeida, P., Jungerman, F., Laranjeira, R., Bressan, R., and Lacerda, A. (2011). Cannabis use before age 15 and subsequent executive functioning. *British Journal of Psychiatry*, 198, 442–447. doi: 10.1192/bjp.bp.110.077479.

Francioli, L. C., , Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., ..., and Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics*, 47(7), 822–826. doi: 10.1038/ng.3292.

Frank, M. C. and Holubar, T. (2015). Replication of Monin, Sawyer, & Marquez (2008, JPSP 95(1), exp. 4). online. Available from: osf.io/pz0my.

Furr, R. M. and Rosenthal, R. (2003). Repeated-measures contrasts for "multiple-pattern" hypotheses. *Psychological Methods*, 8(3), 275–293. doi: 10.1037/1082-989X.8.3.275.

Galliani, E. (2015). Replication report of Fischer, Greitemeyer, and Frey (2008, JPSP, study 2). online. Available from: osf.io/j8bpa.

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

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4), 733–760.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–472. doi: 10.1214/ss/1177011136.

Gelman, A. and Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4), 328–331. doi: 10.1198/000313006x152649.

Genest, C. and Zidek, J. V. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1), 147–148. doi: 10.1214/ss/1177013831.

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach*, volume 20. CRC press.

Gokhale, D. and Press, S. J. (1982). Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution. *Journal of the Royal Statistical Society. Series A (General)*, pages 237–249. doi: 10.2307/2981537.

Goldmann, J. M., Seplyarskiy, V. B., Wong, W. S. W., Vilboux, T., Neerincx, P. B., Bodian, D. L., ..., and Niederhuber, J. E. (2018). Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Obstetrical & Gynecological Survey*, 73(9), 531–532. doi: 10.1097/ogx.0000000000000604.

Goldstein, D. G., Johnson, E. J., and Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *J Consum Res*, 35(3), 440–456. doi: 10.1086/589562.

Goldstein, D. G. and Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, 9(1), 1–14.

Goossens, L., Cooke, R., Hale, A., and Rodić-Wiersma, L. (2008). Fifteen years of expert judgement at TUDelft. *Safety Science*, 46(2), 234–244. doi: 10.1016/j.ssci.2007.03.002.

Gore, S. (1987). Biostatistics and the medical research council. *Medical Research Council News*, 35, 19–20.

Gratten, J., Wray, N. R., Peyrot, W. J., McGrath, J. J., Visscher, P. M., and Goddard, M. E. (2016). Risk of psychiatric illness from advanced paternal age is not predominantly from de novo mutations. *Nature Genetics*, 48(7), 718–724. doi: 10.1038/ng.3577.

Gu, X., Mulder, J., and Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71(2), 229–261. doi: 10.1111/bmsp.12110.

Hallquist, M. (2013). MplusAutomation: Automating Mplus model estimation and interpretation. Package MplusAutomation. Available from: http://cran.r-project.org/web/packages/MplusAutomation/MplusAutomation.pdf.

Hampson, L. V., Whitehead, J., Eleftheriou, D., and Brogan, P. (2014). Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33(24), 4186–4201. doi: 10.1002/sim.6225.

Haran, U. and Moore, D. A. (2010). A simple remedy for overprecision in judgement. 5(7), 467–476. doi: 10.1037/e615882011-200.

Haran, U. and Moore, D. A. (2014). A better way to forecast. *California Management Review*, 57(1), 5–15. doi: 10.1525/cmr.2014.57.1.5.

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

Harms, C. (2018b). *ReplicationBF: Calculating Replication Bayes Factors for Different Scenarios*, R package version 0.0.4.9 edition. Available from: https://github.com/neurotroph/ReplicationBF.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. doi: 10.3102/10769986006002107.

Hedges, L. V. and Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88(2), 359. doi: 10.1037/0033-2909.88.2.359.

Henderson, M. D., de Liver, Y., and Gollwitzer, P. M. (2008). The effects of an implemental mind-set on attitude strength. *Journal of Personality and Social Psychology,*, 94(3), 396–411. doi: 10.1037/0022-3514.94.3.396.

Heron, J., Barker, E. D., Joinson, C., Lewis, G., Hickman, M., Munafò, M., and Macleod, J. (2013). Childhood conduct disorder trajectories, prior risk factors and cannabis use at age 16: birth cohort study. *Addiction*, 108(12), 2129–2138. doi: 10.1111/add.12268.

Hochweber, J. and Hartig, J. (2017). Analyzing organizational growth in repeated cross-sectional designs using multilevel structural equation modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 13(3), 83–97. doi: 10.1027/1614-2241/a000133.

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

Hora, S. C. and Von Winterfeldt, D. (1997). Nuclear waste and future societies: A look into the deep future. *Technological Forecasting and Social Change*, 56(2), 155–170. doi: 10.1016/S0040-1625(97)00075-9.

Hox, J. and Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, 8(2), 157–174. doi: 10.1207/S15328007SEM0802_1.

Hox, J., Moerbeek, M., Kluytmans, A., and Van de Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials. the effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology*, 5, 78. doi: 10.3389/fpsyg.2014.00078.

Hox, J., Van de Schoot, R., and Matthijse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Association*, 6, 87–93.

Hudziak, J. J., van Beijsterveldt, C. E. M., Bartels, M., Rietveld, M. J. H., Rettew, D. C., Derks, E. M., and Boomsma, D. I. (2003). Individual differences in aggression: Genetic analyses by age, gender, and informant in 3-, 7-, and 10-year-old dutch twins. *Behavior Genetics*, 33(5), 575–589. doi: 10.1023/a:1025782918793.

Jacobus, J., Bava, S., Cohen-Zion, M., Mahmood, O., and Tapert, S. (2009). Functional consequences of marijuana use in adolescents. *Pharmacology, Biochemistry and Behavior*, 4, 559–565. doi: 10.1016/j.pbb.2009.04.001.

Janecka, M., Haworth, C. M., Ronald, A., Krapohl, E., Happé, F., Mill, J., Schalkwyk, L. C., Fernandes, C., Reichenberg, A., and Rijsdijk, F. (2017a). Paternal age alters social development in offspring. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(5), 383–390. doi: 10.1016/j.jaac.2017.02.006.

Janecka, M., Rijsdijk, F., Rai, D., Modabbernia, A., and Reichenberg, A. (2017b). Advantageous developmental outcomes of advancing paternal age. *Translational Psychiatry*, 7(6), e1156. doi: 10.1038/tp.2017.125.

Janiszewski, C. and Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science*, 19(2), 121–127. doi: 10.1111/j.1467-9280.2008.02057.x.

JASP Team (2016). Jasp (version 0.8.0)[computer software]. Available from: https://jasp-stats.org/.

JASP Team (2018). Jasp (version 0.9.0)[computer software]. Available from: https://jasp-stats.org/.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, Oxford, 3 edition.

Jenkins, T. G., Aston, K. I., Pflueger, C., Cairns, B. R., and Carrell, D. T. (2014). Age-associated sperm DNA methylation alterations: Possible implications in offspring disease susceptibility. *PLoS Genetics*, 10(7), e1004458. doi: 10.1371/journal.pgen.1004458.

Jensen, P. S., Rubio-Stipec, M., Canino, G., Bird, H. R., Dulcan, M. K., Schwab-Stone, M. E., and Lahey, B. B. (1999). Parent and child contributions to diagnosis of mental disorder: are both informants always necessary? *Journal of the American Academy of Child & Adolescent Psychiatry*, 38(12), 1569–1579. doi: 10.1097/00004583-199912000-00019.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., and Feldman, B. M. (2010a). Methods to elicit beliefs for bayesian priors: a systematic review. *Journal of Clinical Epidemiology*, 63(4), 355–369. doi: 10.1016/j.jclinepi.2009.06.003.

Johnson, S. R., Tomlinson, G. A., Hawker, G. A., Granton, J. T., Grosbein, H. A., and Feldman, B. M. (2010b). A valid and reliable belief elicitation method for Bayesian priors. *Journal of Clinical Epidemiology*, 63(4), 370–383. doi: 10.1016/j.jclinepi.2009.08.005.

Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., ..., and Stefansson, K. (2017). Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 549(7673), 519–522. doi: 10.1038/nature24018.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Publications.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. doi: 10.1080/01621459.1995.10476572.

Khandwala, Y. S., Baker, V. L., Shaw, G. M., Stevenson, D. K., Lu, Y., and Eisenberg, M. L. (2018). Association of paternal age with perinatal outcomes between 2007 and 2016 in the united states: population based cohort study. *BMJ*, page k4372. doi: 10.1136/bmj.k4372.

Kiernan, K. E. (1997). Becoming a young parent: A longitudinal study of associated factors. *The British Journal of Sociology*, 48(3), 406. doi: 10.2307/591138.

Kline, R. B. et al. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. American Psychological Association. doi: 10.1037/10693-000.

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., ..., and Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412), 471–475. doi: 10.1038/nature11396.

Kraan, B. C. P. (2002). Probabilistic inversion in uncertainty analysis: and related topics.

Kruschke, J. K. (2011). Introduction to special section on Bayesian data analysis. *Perspectives on Psychological Science*, 6(3), 272–273. doi: 10.1177/1745691611406926.

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, San Diego, CA, 2nd edition.

Kuiper, R. M., Buskens, V., Raub, W., and Hoijtink, H. (2012). Combining statistical evidence from several studies: A method using bayesian updating and an example from research on trust problems in social and economic exchange. *Sociological Methods & Research*, 42(1), 60–81. doi: 10.1177/0049124112464867.

Kuiper, R. M. and Hoijtink, H. (2013). A Fortran 90 program for the generalization of the order-restricted information criterion. *Journal of Statistical Software*, 54, 1–19. doi: 10.18637/jss.v054.i08.

Lane, K. A. and Gazerian, D. (2016). Replication of Henderson, de Liver, & Gollwitzer (2008, JPSP, expt. 5). Available from: osf.io/79dey.

Lawlor, D. A., Mortensen, L., and Andersen, A.-M. N. (2011). Mechanisms underlying the associations of maternal age with adverse perinatal outcomes: A sibling study of 264695 Danish women and their firstborn offspring. *International Journal of Epidemiology*, 40(5), 1205–1214. doi: 10.1093/ije/dyr084.

Ledgerwood, A. (2014). Introduction to the special section on advancing our methods and practices. *Perspectives on Psychological Science*, 9(3), 275–277. doi: 10.1177/1745691613513470.

Lee, B. K. and McGrath, J. J. (2015). Advancing parental age and autism: multifactorial pathways. *Trends in Molecular Medicine*, 21(2), 118–125. doi: 10.1016/j.molmed.2014.11.005.

Lee, S.-Y. and Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4), 653–686. doi: 10.1207/s15327906mbr3904_4.

Leydold, J. and Hörmann, W. (2015). *Runuran: R Interface to the UNU.RAN Random Variate Generators*. R package version 0.23.0. Available from: http://CRAN.R-project.org/package=Runuran.

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

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Lüdtke, O., Marsh, H. W., Robitzsch, A., and Trautwein, U. (2011). A 2× 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, 16(4), 444–467. doi: 10.1037/a0024376.

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

Lynch, S. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York, NY. doi: 10.1007/978-0-387-71265-9.

Maas, C. J. and Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1(3), 86–92. doi: 10.1027/1614-1881.1.3.86.

Mahmood, O. M., Jacobus, J., Bava, S., Scarlett, A., and Tapert, S. F. (2010). Learning and memory performances in adolescent users of alcohol and marijuana: Interactive effects. *Journal of Studies on Alcohol and Drugs*, 71(6), 885–894. doi: 10.15288/jsad.2010.71.885.

Mahu, I. T., Doucet, C., O'Leary-Barrett, M., and Conrod, P. J. (2015). Can cannabis use be prevented by targeting personality risk in schools? Twenty-four-month outcome of the adventure trial on cannabis use: a cluster-randomized controlled trial. *Addiction*, 110(10), 1625–1633. doi: 10.1111/add.12991.

Mäkelä, P. and Huhtanen, P. (2010). The effect of survey sampling frame on coverage: The level of and changes in alcohol-related mortality in Finland as a test case. *Addiction*, 105(11), 1935–1941. doi: 10.1111/j.1360-0443.2010.03069.x.

Manning, W. D., Trella, D., Lyons, H., and Toit, N. C. D. (2010). Marriageable women: A focus on participants in a community healthy marriage program. *Family Relations*, 59(1), 87–102. doi: 10.1111/j.1741-3729.2009.00588.x.

Marsman, M., Schönbrodt, F. D., Morey, R. D., Yao, Y., Gelman, A., and Wagenmakers, E.-J. (2017). A Bayesian bird's eye view of 'replications of important results in social psychology'. *Royal Society Open Science*, 4(1), 160426. doi: 10.1098/rsos.160426.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163. doi: 10.1037/1082-989X.9.2.147.

McCabe, S. E., Kloska, D. D., Veliz, P., Jager, J., and Schulenberg, J. E. (2016). Developmental course of non-medical use of prescription drugs from adolescence to adulthood in the united states: national longitudinal data. *Addiction*, 111(12), 2166–2176. doi: 10.1111/add.13504.

McGrath, J. J., Petersen, L., Agerbo, E., Mors, O., Mortensen, P. B., and Pedersen, C. B. (2014). A comprehensive assessment of parental age and psychiatric disorders. *JAMA Psychiatry*, 71(3), 301. doi: 10.1001/jamapsychiatry.2013.4081.

McMahon, C. A., Gibson, F. L., Allen, J. L., and Saunders, D. (2007). Psychosocial adjustment during pregnancy for older couples conceiving through assisted reproductive technology. *Human Reproduction*, 22(4), 1168–1174. doi: 10.1093/humrep/del502.

McNeish, D. (2016a). On using bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5), 750–773. doi: 10.1080/10705511.2016.1186549.

McNeish, D. M. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics*, 41(1), 27–56. doi: 10.3102/1076998615621299.

Menezes, P. R., Lewis, G., Rasmussen, F., Zammit, S., Sipos, A., Harrison, G. L., Tynelius, P., and Gunnell, D. (2010). Paternal and maternal ages at conception and

risk of bipolar affective disorder in their offspring. *Psychological Medicine*, 40(3), 477–485. doi: 10.1017/s003329170999064x.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22(3), 1142–1160. doi: 10.1214/aos/1176325622.

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

Meuleman, B. and Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods*, 3, 45–58.

Monin, B., Sawyer, P. J., and Marquez, M. J. (2008). The rejection of moral rebels: Resenting those who do the right thing. *Journal of Personality and Social Psychology*, 95(1), 76–93. doi: 10.1037/0022-3514.95.1.76.

Morales Nápoles, O. (2010). *Bayesian belief nets and vines in aviation safety and other applications*. PhD thesis.

Morey, Richard D Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*, R package version 0.9.12-4.2 edition. Available from: https://CRAN.R-project.org/package=BayesFactor.

Morgan, M. G., Henrion, M., and Small, M. (1992). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge university press. doi: 10.1017/cbo9780511840609.

Morris, D. E., Oakley, J. E., and Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, 52, 1–4. doi: 10.1016/j.envsoft.2013.10.010.

Muthén, B. O. and Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, 2, 371–402. doi: 10.1037/1082-989X.2.4.371.

Muthén, L. K. and Muthén, B. O. (1998-2012). *Mplus user's guide*. Muthén & Muthén, Los Angeles, CA, 7th edition.

Muthén, L. K. and Muthén, B. O. (1998-2017). *Mplus user's guide*. Muthén & Muthén, Los Angeles, CA, 8th edition.

Muthén, L. K. and 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. doi: 10.1207/S15328007SEM0904_8.

Myrskylä, M., Barclay, K., and Goisis, A. (2017). Advantages of later motherhood. *Der Gynäkologe*, 50(10), 767–772. doi: 10.1007/s00129-017-4124-1.

Myrskylä, M. and Fenelon, A. (2012). Maternal age and offspring adult health: Evidence from the health and retirement study. *Demography*, 49(4), 1231–1257. doi: 10.1007/s13524-012-0132-x.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, 14(9), 1105–1107. doi: 10.1038/nn.2886.

Ntzoufras (2009). *Models for Positive Continuous Data, Count Data, and Other GLM-Based Extensions*, chapter 8, pages 275–304. Wiley-Blackwell.

Nybo Andersen, A.-M. and Urhoj, S. K. (2017). Is advanced paternal age a health risk for the offspring? *Fertility and Sterility*, 107(2), 312–318. doi: 10.1016/j.fertnstert.2016.12.019.

Oakley, J. E. and O'Hagan, A. (2010). *SHELF: The Sheffield elicitation framework (Version 2.0) [Computer software and documentation]*. School of Mathematics and Statistics, University of Sheffield. Available from: www.tonyohagan.co.uk/shelf.

Oakley, J. E. and O'Hagan, A. (2016). *SHELF: the Sheffield Elicitation Framework (version 3.0) [Computer software and documentation]*. School of Mathematics and Statistics, University of Sheffield, UK. Available from: http://tonyohagan.co.uk/shelf.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons. doi: 10.1002/0470033312.

Okada, K. (2013). Is omega squared less biased? a comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, 40(2), 129–147. doi: 10.2333/bhmk.40.129.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660. doi: 10.1177/1745691612462588.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251). doi: 10.1126/science.aac4716.

Orlebeke, J. F., Knol, D. L., Boomsma, D. I., and Verhulst, F. C. (1998). Frequency of parental report of problem behavior in children decreases with increasing maternal age at delivery. *Psychological Reports*, 82(2), 395–404. doi: 10.2466/pr0.1998.82.2.395.

Pashler, H. and Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6), 528–530. doi: 10.1177/1745691612465253.

Patil, P., Peng, R. D., and 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. doi: 10.1177/1745691616646366.

Peeters, M., Monshouwer, K., Janssen, T., Wiers, R. W., and Vollebergh, W. A. (2014). Working memory and alcohol use in at-risk adolescents: A 2-year follow-up. *Alcoholism: Clinical and Experimental Research*, 38(4), 1176–1183. doi: 10.1111/acer.12339.

Petrides, M. and Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, 20, 249–262. doi: 10.1016/0028-3932(82)90100-2.

Pierce, R. (2014). Correlation. Available from: http://www.mathsisfun.com/data/correlation.html.

Plummer, M. (2013). *JAGS Version 3.4.0 user manual computing [Computer software manual]*. Available from: http://mcmc-jags.sourceforge.net.

Plummer, M. (2015). *rjags: Bayesian Graphical Models using MCMC*. R package version 3-15. Available from: http://CRAN.R-project.org/package=rjags.

Press, S. J. (2009). *Subjective and objective Bayesian statistics: principles, models, and applications*, volume 590. John Wiley & Sons.

Prifitera, A. and Saklofske, D. H. (1998). *WISC-III clinical use and interpretation: Scientist-practitioner perspectives*. Elsevier.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 3.0.1 edition. Available from: `https://www.R-project.org/`.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 3.2.0 edition. Available from: `https://www.R-project.org/`.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 3.3.0 edition. Available from: `https://www.R-project.org/`.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 3.4.0 edition.

Renk, K. and Phares, V. (2004). Cross-informant ratings of social competence in children and adolescents. *Clinical Psychology Review*, 24(2), 239–254. doi: `10.1016/j.cpr.2004.01.004`.

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., ..., and Verhulst, F. C. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child & Adolescent Psychology*, 42(2), 262–273. doi: `10.1080/15374416.2012.717870`.

Reynolds, W. M. (2000). *Reynolds Adolescent Depression Scale*. doi: `10.1002/9780470479216.corpsy0798`.

Robson, K. and Pevalin, D. J. (2007). Gender differences in the predictors and socio-economic outcomes of young parenthood in Great Britain. *Research in Social Stratification and Mobility*, 25(3), 205–218. doi: `10.1016/j.rssm.2007.08.002`.

Rocchetti, M., Crescini, A., Borgwardt, S., Caverzasi, E., Politi, P., Atakan, Z., and Fusar-Poli, P. (2013). Is cannabis neurotoxic for the healthy brain? A meta-analytical review of structural brain alterations in non-psychotic users. *Psychiatry and Clinical Neurosciences*, 67(7), 483–492. doi: `10.1111/pcn.12085`.

Roe-Sepowitz, D. E. (2009). Comparing male and female juveniles charged with homicide child maltreatment, substance abuse, and crime details. *Journal of Interpersonal Violence*, 24(4), 601–617. doi: `10.1177/0886260508317201`.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641. doi: `10.1037/0033-2909.86.3.638`.

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

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, 57(5), 416–428. doi: `10.1037/h0042040`.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. 12(4), 1151–1172. doi: `10.1214/aos/1176346785`.

Rupp, A. A., Dey, D. K., and 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. doi: 10.1207/s15328007sem1103_7.

Saha, S., Barnett, A. G., Buka, S. L., and McGrath, J. J. (2009). Maternal age and paternal age are associated with distinct childhood behavioural outcomes in a general population birth cohort. *Schizophrenia Research*, 115(2-3), 130–135. doi: 10.1016/j.schres.2009.09.012.

Sandberg-Thoma, S. E. and Dush, C. M. K. (2014). Indicators of adolescent depression and relationship progression in emerging adulthood. *Journal of Marriage and Family*, 76(1), 191–206. doi: 10.1111/jomf.12081.

Sandin, S., Hultman, C. M., Kolevzon, A., Gross, R., MacCabe, J. H., and Reichenberg, A. (2012). Advancing maternal age is associated with increasing risk for autism: A review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(5), 477–486.e1. doi: 10.1016/j.jaac.2012.02.018.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147–177. doi: 10.1037//1082-989x.7.2.147.

Scharkow, M., Festl, R., and Quandt, T. (2014). Longitudinal patterns of problematic computer game use among adolescents and adults: a 2-year panel study. *Addiction*, 109(11), 1910–1917. doi: 10.1111/add.12662.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13(2), 90–100. doi: 10.1037/a0015108.

Schweinsburg, A., Brown, S., and Tapert, S. (2008). The influence of marijuana use on neurocognitive functioning in adolescents. *Current Drug Abuse Reviews*, 1, 99–111. doi: 10.2174/1874473710801010099.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. doi: 10.1037//0033-2909.86.2.420.

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

Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. doi: 10.1177/0956797614567341.

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

Spiegelhalter, D. J., Myles, J. P., Jones, D. R., and Abrams, K. R. (2000). Bayesian methods in health technology assessment: A review. *Health Technology Assessment*, 4(38), 1–130.

Stan Development Team (2014). *Stan: A C++ library for probability and sampling (Version 2.11.0) [Software]*. Available from: http://mc-stan.org.

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

Steiger, J. H. (2004). Beyond the f test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9(2), 164–182. doi: 10.1037/1082-989X.9.2.164.

Stokes, J. M., Pogge, D. L., and 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. doi: 10.1080/00223891.2012.730084.

Te Velde, E. R. (2002). The variability of female reproductive ageing. *Human Reproduction Update*, 8(2), 141–154. doi: 10.1093/humupd/8.2.141.

Tearne, J. E., Robinson, M., Jacoby, P., Allen, K. L., Cunningham, N. K., Li, J., and McLean, N. J. (2016). Older maternal age is associated with depression, anxiety, and stress symptoms in young adult female offspring. *Journal of Abnormal Psychology*, 125(1), 1–10. doi: 10.1037/abn0000119.

Tearne, J. E., Robinson, M., Jacoby, P., Li, J., Newnham, J., and McLean, N. (2015). Does late childbearing increase the risk for behavioural problems in children? A longitudinal cohort study. *Paediatric and Perinatal Epidemiology*, 29(1), 41–49. doi: 10.1111/ppe.12165.

Tolvanen, A. (2000). *Latenttien kasvukäyrä- ja simplex-mallien teoriaa ja sovelluksia pitkittäisaineistoissa kehityksen ja muutoksen analysointiin [Latent growth and simplex models: Theory and applications in longitudinal models for analysis of development and change]*. Department of Statistics, University of Jyväskylä.

Trillingsgaard, T. and Sommer, D. (2018). Associations between older maternal age, use of sanctions, and children's socio-emotional development through 7, 11, and 15 years. *European Journal of Developmental Psychology*, 15(2), 141–155. doi: 10.1080/17405629.2016.1266248.

Turlach, B. A. and Weingessel, A. (2013). *quadprog: Functions to solve Quadratic Programming Problems*, r package version 1.5-5 edition. Available from: https://CRAN.R-project.org/package=quadprog.

Van Aert, R. C. (2018). *puniform: Meta-Analysis Methods Correcting for Publication Bias*, R package version 0.1.0 edition. Available from: https://CRAN.R-project.org/package=puniform.

Van Aert, R. C. and Van Assen, M. A. (2017a). Bayesian evaluation of effect size after replicating an original study. *PloS One*, 12(4), e0175302. doi: 10.1371/journal.pone.0175302.

Van Aert, R. C. and Van Assen, M. A. (2017b). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. *Behavior Research Methods*, pages 1–25. doi: 10.3758/s13428-017-0967-6.

Van Buuren, S. (2012). *Flexible imputation of missing data*. Chapman & Hall/CRC. doi: 10.1201/b11826.

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

Van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., and Van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, 6, 25216. doi: 10.3402/ejpt.v6.25216.

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

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., and Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods*, 22, 217–239. doi: 10.1037/met0000100.

Van der Lee, J. H., Wesseling, J., Tanck, M. W. T., and Offringa, M. (2008). Efficient ways exist to obtain the optimal sample size in clinical trials in rare diseases. *Journal of Clinical Epidemiology*, 61(4), 324–330. doi: 10.1016/j.jclinepi.2007.07.008.

Van Erp, S., Mulder, J., and Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods*, 23(2), 363–388. doi: 10.1037/met0000162.

Van Lenthe, J. (1993). A blueprint of ELI: A new method for eliciting subjective probability distributions. *Behavior Research Methods, Instruments, & Computers*, 25(4), 425–433. doi: 10.3758/BF03204541.

Veen, D., Stoel, D., Zondervan-Zwijnenburg, M., and van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgment. *Frontiers in Psychology*, 8(2110). doi: 10.3389/fpsyg.2017.02110.

Verhagen, J. and Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143(4), 1457–1475. doi: 10.1037/a0036731.

Wabersich, D. and Vandekerckhove, J. (2013). Extending JAGS: A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behavioral Resarch Methods*, 46(1), 15–28. doi: 10.3758/s13428-013-0369-3.

Walker, E. and Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of general internal medicine*, 26(2), 192–196. doi: 10.1007/s11606-010-1513-8.

Wechsler, D. (1991). *Manual for the Wechsler intelligence scale for children - Third Eediction (WISC-III)*. Psychological Corporation, San Antonio, TX.

Weiser, M., Reichenberg, A., Werbeloff, N., Kleinhaus, K., Lubin, G., Shmushkevitch, M., ..., and Davidson, M. (2008). Advanced parental age at birth is associated with poorer social functioning in adolescent males: Shedding light on a core symptom of schizophrenia and autism. *Schizophrenia Bulletin*, 34(6), 1042–1046. doi: 10.1093/schbul/sbn109.

Weiss, L. G., Saklofske, D. H., Prifitera, A., and Holdnack, J. A. (2006). *WISC-IV advanced clinical interpretation*. Academic Press.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., and Morales-Nápoles, O. (2017). Expert judgement for dependence in probabilistic modelling: a systematic literature review and future research directions. *European Journal of Operational Research*, 258(3), 801–819. doi: 10.1016/j.ejor.2016.10.018.

Winkler (1968). The consensus of subjective probability distributions. *Management Science*, 15(2), 61–75. doi: 10.1287/mnsc.15.2.b61.

Wong, W. S. W., Solomon, B. D., Bodian, D. L., Kothiyal, P., Eley, G., Huddleston, K. C., Baker, R., ..., and Niederhuber, J. E. (2016). New observations on maternal age effect on germline de novo mutations. *Nature Communications*, 7(1). doi: 10.1038/ncomms10486.

Yuan, K.-H. and Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, 30(2), 141–167. doi: 10.3102/10769986030002141.

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., and Van de Schoot, R. (2017a). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development*, 14(4), 305–320. doi: 10.1080/15427609.2017.1370966.

Zondervan-Zwijnenburg, M., Van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., and Van de Schoot, R. (2017b). Application and evaluation of an expert judgment elicitation procedure for correlations. *Frontiers in Psychology*, 8, 90. doi: 10.3389/fpsyg.2017.00090.

Zondervan-Zwijnenburg, M. A. J. (2018). *ANOVAreplication: Test ANOVA Replications by Means of the Prior Predictive p-Value*. R package version 1.1.3. Available from: https://CRAN.R-project.org/package=ANOVAreplication.

Zondervan-Zwijnenburg, M. A. J. (2019). *Replication: Test Replications by Means of the Prior Predictive p-Value*. R package version 0.1.0. Available from: https://CRAN.R-project.org/package=Replication.

Zondervan-Zwijnenburg, M. A. J., Van de Schoot, R., and Hoijtink, H. Testing ANOVA replications by means of the prior predictive *p*-value [online]. (2019). doi: 10.31234/osf.io/6myqh.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological methods*, 12(4), 399. doi: 10.1037/1082-989X.12.4.399.

Zweers, I. (2018). *"Shape sorting" students for special education services?* PhD thesis, Utrecht University.

Samenvatting

Voorkennis is kennis die men heeft buiten de huidige data om. Deze kennis kan bijvoorbeeld gebaseerd zijn op eerder onderzoek of op ervaring. In dit proefschrift wordt duidelijk hoe deze kennis verzameld kan worden, hoe deze omgezet kan worden in een prior verdeling en hoe deze kennis geëvalueerd kan worden. Voordat de verschillende hoofdstukken worden samengevat, worden twee manieren om voorkennis te gebruiken in analyses kort beschreven.

Informatieve Prior Verdeling

Informatieve prior verdelingen zijn onderdeel van Bayesiaanse statistiek. Bayesiaanse statistiek drukt uitkomsten uit in kansen. Een parameter zoals het gemiddelde wordt gebaseerd op zowel data als op onze verwachting omtrent het gemiddelde. De verwachting wordt uitgedrukt in een kansverdeling. Als het gaat om het gemiddelde van IQ (i.e., μ_{IQ}), dan kan onze verwachting zijn dat het gemiddelde 100 is, maar aangezien we niet 100% zeker zijn dat het gemiddelde in deze dataset 100 is, nemen we ook onzekerheid mee. We drukken dat bijvoorbeeld uit in een normaalverdeling met een gemiddelde van 100, en een standaarddeviatie van 10. Dat is:

$$\mu_{\text{IQ}} \sim N(100, 10). \tag{9.1}$$

Deze kansverdeling wordt ook weergeven in Figuur 1.1. Wanneer de data gecombineerd wordt met de prior verwachting, is de uitkomst ook een kansverdeling: de posterior verdeling. De posterior kansverdeling geeft aan wat de kans is van elke waarde om het gemiddelde van IQ te zijn.

In Bayesiaanse statistiek wordt voorkennis dus onderdeel van de analyse, net zo goed als de data. Wanneer er weinig tot geen voorkennis is, wordt dit uitgedrukt in een zeer vlakke prior verdeling die een wijde reeks waarden een soortgelijke kans toebedeeld. Deze dissertatie toont aan in Hoofdstuk 3 en 4 hoe voorkennis verkregen kan worden en hoe deze omgezet kan worden in prior verdelingen en wat de impact is ervan is. Daarnaast wordt de voorkennis ook op verschillende manieren geëvalueerd: met een simulatie studie in Hoofdstuk 2, met een robuustheidsanalyse in Hoofdstuk 3, met visualisaties in Hoofdstuk 4, en met de 'prior predictieve p-waarde' Box (1980) in Hoofdstuk 5, 6 en 7.

Informatieve Hypothese

Ook informatieve hypotheses zijn een manier om voorkennis te includeren. De informatieve hypothese bevat beperkingen voor de waarden die een parameter aan kan nemen. Zo zijn er beperkingen in de reeks waarden die een parameter aan kan nemen (bijv. $\mu_{\text{IQ}} > 80$), beperkingen in de volgorde van grootheid tussen parameters (bijv. $\mu_{\text{IQ-regulier basisonderwijs}} > \mu_{\text{IQ-speciaal basisonderwijs}}$), en gelijkheidsbeperkingen die stellen dat de parameter gelijk moet zijn aan één specifiek getal (bijv. $\mu_{\text{IQ}} = 100$) of aan een andere parameter (bijv. $\mu_{\text{IQ-Utrecht}} = \mu_{\text{IQ-Leiden}}$).

Om informatieve hypotheses te testen kan (1) een frequentistische p-waarde worden berekend (zie Silvapulle and Sen, 2005), (2) een informatie criterium worden berekend voor de vergelijking van verschillende informatieve hypotheses (zie Kuiper and Hoijtink, 2013), of (3) een Bayes factor worden berekend die verschillende hypotheses vergelijkt (zie Gu et al., 2018).

De Bayes factor methode wordt toegepast in Hoofdstuk 8. Daarnaast vormen informatieve hypotheses een onderdeel van de 'prior predictive check' in de Hoofdstukken 5, 6 en 7.

Het Verkrijgen en Formaliseren van Voorkennis

In Hoofdstuk 3 wordt een systematische literatuurstudie uitgevoerd om voorkennis te verzamelen met betrekking tot een complex longitudinaal analyse-model. Helaas zijn er geen bruikbare meta-analyses die data aanleveren waarop prior informatie gebaseerd kan worden. Meerdere reviews impliceren dat het hoofdeffect klein zal zijn, maar mogelijk wel aanwezig. De review resultaten geven een richting aan voor de prior verdelingen, maar meer informatie is nodig. Vervolgens wordt er in de literatuur gezocht naar studies die gemiddelden presenteren voor het gebruikte meetinstrument. Twee experts, een gedragspsycholoog en een professor ontwikkelingspsychopathologie, beoordelen vervolgens in hoeverre de gevonden gemiddelden in de studies van toepassing zijn voor de onderzoeksgroepen. De bruikbare informatie blijkt beperkt, maar kan opnieuw wel richting geven voor de prior verdelingen. In combinatie met algemene kennis over ontwikkeling en de meetinstrumenten kunnen vrij algemene prior verdelingen opgesteld worden. De winst die volgens Hoofdstuk 2 behaald kan worden met zeer informatieve priors heeft een prijs: het zoeken naar voorkennis is niet altijd gemakkelijk en levert ook niet altijd het gewenste resultaat op. De inhoudelijk onderzoeker kan het beste inschatten wanneer het zoeken naar voorkennis de moeite waard zal zijn. Een voorbeeld is (Zweers, 2018, Chapter 5). Zij maakten gebruik van de richtlijnen gepresenteerd in Zondervan-Zwijnenburg et al. (2017a) en vond daarmee de mogelijkheid om succesvolle analyses te verrichten.

In hoofdstuk 4 wordt een procedure ontwikkeld om inhoudelijke experts te bevragen op een correlatie. In dit geval werden ontwikkelingspsychologen bevraagd op hun kennis van de relatie tussen schoolvoortgang en IQ bij kinderen met en zonder autisme spectrum stoornis (ASS) die het middelbaar speciaal onderwijs voor kinderen met

ernstige gedragsproblemen bezochten. De methode beslaat het toewijzen van kansen met behulp van stickers aan mogelijke uitkomsten voor de correlatie.

De indruksvaliditeit van het ontwikkelde instrument was goed. De convergente validiteit die de samenhang met andere correlatie-schattingen uitdrukt was .42 en .59 voor respectievelijk de groep met en zonder ASS. De onzekerheid rond deze schattingen was dermate hoog dat er geen sterke conclusies aan verbonden konden worden. Het absolute verschil in de schatting ten opzichte van de andere correlatiemaat was klein. De puntschattingen uit een alternatieve correlatie-vraag lagen allen binnen de geschatte verdelingen van de experts. Naast validiteit werd ook betrouwbaarheid onderzocht. De gedragspsychologen kregen de test ruim vier maanden later nog eens opgestuurd om een indicatie van test-hertestbetrouwbaarheid te verkrijgen. Eén psycholoog vond niet de gelegenheid om de vragenlijst in te vullen. De resultaten van de anderen wezen op een onvoldoende test-hertestbetrouwbaarheid. Een mogelijke conclusie is dat het voor sommige expertgroepen zeer belangrijk is om hen rustig de gelegenheid te geven om de vragen te beantwoorden, bij voorkeur in de aanwezigheid van de onderzoeker zodat ook vragen beantwoord kunnen worden. Hoofdstuk 4 beschrijft verder hoe de verkregen resultaten van de experts omgezet kunnen worden in bruikbare prior verdelingen. De expertkennis wordt in deze stap geformaliseerd. Vervolgens wordt de kennis geüpdated met data en worden de uiteindelijke resultaten geëvalueerd.

Het Evalueren van Voorkennis

Bayesiaanse priors

In hoofdstuk 2 wordt de invloed van het includeren van voorkennis in analyses verhelderd door middel van een simulatiestudie. Het uitgangspunt is een latent groeimodel voor twee groepen: een referentiegroep met 50 tot 10.000 participanten en een speciale groep met 5 tot 50 participanten. Er is dus één zeer kleine groep en een tweede groep die meer individujën kan omvatten. De focus ligt op het verschil in lineaire groei tussen de twee groepen vastgesteld op de eerste meting. Frequentistische analyses met een Maximum likelihood (ML) schatter en Bayesiaanse analyses met weinig voorkennis laten zien dat de schattingen ook met zeer kleine groepen dicht bij de populatiewaarde liggen. Tegelijkertijd resulteren de analyses met de ML schatter vaak in waarschuwingen over onrealistische schattingen en daarmee onbruikbare resultaten. In de Bayesiaanse analyses komen deze schattingen niet voor doordat de prior distributies geen onrealistische schattingen zoals negatieve varianties toestaan. Een problematisch resultaat voor de ML schatter en de Bayesiaanse analyses met weinig voorkennis is de statistische power om een nul-effect te verwerpen: deze is zeer laag. Door het gebrek aan informatie in de data, kunnen er weinig conclusies aan de resultaten verbonden worden. Pas wanneer er sterke prior verwachtingen geformuleerd kunnen worden, zijn de resultaten specifiek genoeg om er conclusies aan te verbinden. Met name voorkennis ten opzichte van de kleine groep is hierin relevant. Dat het juist voor deze groep lastig is om voorkennis te vinden, wordt ook duidelijk in Hoofdstuk 3.

Replicatie

Bij een replicatie poging spelen twee studies een rol: de originele studie en de nieuwe studie. De originele studie kan gezien worden als voorkennis voor een nieuwe studie. We zouden kunnen stellen dat de nieuwe studie in lijn moet zijn met de voorkennis (dat is de originele studie) om de studie gerepliceerd te noemen. Om deze evaluatie uit te voeren is er de zogenaamde 'prior predictive check' van Box (1980). In de test worden datasets aangemaakt die te verwachten zijn gegeven de aanwezige voorkennis. Vervolgens wordt een toetsingsgrootheid gekozen die voor elke voorspelde dataset uitgerekend wordt. De toetsingsgrootheid wordt dan ook voor de nieuwe studie berekend. Wanneer de nieuwe studie een extreme score behaalt ten opzichte van de voorspelde studies, verwerpen we de replicatie. Een extreme score is dan bijvoorbeeld een score die in de uiterste vijf procent van de voorspelde studies ligt. Wanneer een nieuwe studie niet in de extremen scoort, verwerpen we replicatie niet.

Als toetsingsgrootheid kiezen wij voor $\bar{F}$ waarmee een informatieve hypothese, H_0, geëvalueerd wordt. De informatieve hypothese is bijvoorbeeld $\mu_1 > \mu_2 > \mu_3$ in geval van een ANOVA studie. De inhoud van de informatieve hypothese wordt gebaseerd op de claims van de originele studie. Door willekeurige variatie kan ook een studie die te verwachten gegeven het origineel deenige afwijking vertonen ten opzichte van de replicatie hypothese H_0. De vraag is dus niet of de nieuwe studie H_0 exact repliceert, maar of de nieuwe studie niet meer van die hypothese afwijkt dan datasets die je zou verwachten gegeven de originele studie. Hoofdstuk 5 geeft een uitleg van de methode voor het ANOVA model. Daarbij wordt ook een online interactieve R Shiny omgeving gepresenteerd en het R-package `ANOVAreplication` (Zondervan-Zwijnenburg, 2018) waarmee de gebruiker de toets eenvoudig zelf uit kan voeren. Hoofdstuk 6 legt uit hoe de methode toegepast kan worden op structurele vergelijkingsmodellen met behulp van het `Replication` R-package (Zondervan-Zwijnenburg, 2019). In Hoofdstuk 7 worden meerdere replicatie-vragen en methoden naast elkaar gezet in de context van kleine steekproeven. Bayes factor methoden hebben het voordeel met kleine steekproeven dat ze geen probleem kennen met statistische power. De prior predictieve p-waarde heeft als voordeel dat complexe hypotheses en modellen in een enkele toets opgenomen kunnen worden. Uiteindelijk hangt de keuze van onderzoeksmethode altijd samen met de onderzoeksvraag. Voor alle methoden geldt: hoe groter de steekproeven, hoe groter de betekenis van de uitkomst.

Cross-validatie

Tenslotte bevat deze dissertatie een multi-cohort studie met data van het Nederlandse Tweelingen Register (NTR), TRacking Adolescents' Individual Lives Survey (TRAILS), Generation R (Gen-R) en Research on Adolescent Development and Relationships-Young cohort (RADAR-Y). De onderzoeksvraag in deze studie is: Wat is de relatie tussen de leeftijd waarop ouders kinderen krijgen en de psychosociale klachten van het kind op 10-jarige leeftijd? Voor de beantwoording van deze onderzoeksvraag wordt gebruik gemaakt van een vorm van cross-validatie met een exploratieve en

confirmatieve dataset voor elk cohort. Op basis van exploratieve resultaten worden concurrerende informatieve hypotheses opgesteld. Vervolgens wordt met Bayes factoren in de confirmatieve datasets geëvalueerd in hoeverre de data ondersteuning biedt voor elk van de hypotheses. Daarnaast wordt de relatieve steun voor elk van de hypotheses ten opzichte van elkaar geëvalueerd en uitgedrukt in zogenoemde 'posterior model probabilities'. Uiteindelijk wordt de relatieve ondersteuning bij elkaar genomen om te evalueren hoe de relatieve steun is door alle cohorten samen. Dat is: welke hypothese krijgt de meeste steun onafhankelijk van specifieke cohort kenmerken en meetmethoden?

De conclusie van de studie is dat de leeftijd waarop ouders kinderen krijgen geen invloed heeft op internaliserende problematiek op 10-jarige leeftijd. Met betrekking tot externaliserende problemen, wordt er wel een negatief verband gevonden tussen de leeftijd van moeders en externaliserende problemen zoals gerapporteerd door beide ouders: moeders die ouder zijn op het moment dat hun kind geboren wordt, rapporteren minder vaak externaliserend probleemgedrag. Wanneer sociaal-economische status en sekse van het kind worden opgenomen in het model, verdwijnt het verband tussen de leeftijd van moeder en de rapportage van probleemgedrag door vaders. Voor probleemgedrag van het kind volgens moeders wordt er na correctie voor leeftijd en sekse evenveel ondersteuning gevonden voor de hypothese van geen effect, als voor de hypothese van een lineair negatief effect. De moeder-rapportages van probleemgedrag hangen ook negatief samen met de leeftijd van vader, maar dit verband verdwijnt wanneer sociaal-economische status en sekse worden opgenomen in het model.

Door gebruik te maken van een exploratieve en confirmatieve dataset, wordt het gevaar kleiner dat we afwijkingen in de steekproef interpreteren als populatie-effecten. Ook het samenvoegen van ondersteuning van hypotheses vanuit verschillende datasets vermindert de kans dat we onechte effecten interpreteren. Het samenvoegen van verschillende datasets is vaak lastig doordat de meetinstrumenten verschillen. Door de resultaten samen te vatten in informatieve hypotheses die op alle meetinstrumenten van toepassing waren, kon deze uitdaging eenvoudig aangepakt worden.

Discussie

Onderzoek maakt de meeste vooruitgang wanneer zij gebruik maakt van de resultaten en inzichten van anderen. Vandaar ook het Google Scholar motto: "Staan op de schouders van reuzen". Informatieve prior verdelingen en informatieve hypotheses zijn manieren om voorkennis te formaliseren en evalueren.

Het doel van deze dissertatie was om te laten zien hoe voorkennis geformaliseerd en geëvalueerd kan worden. In Deel I lag de nadruk op het formaliseren van voorkennis. Hoofdstuk 2 liet zien hoe voorkennis bevorderend kan werken met het oog op een aantal statistische eigenschappen. Vooral voor de kleinere groep in een studie kan voorkennis voordelig werken. Hoofdstuk 3 echter, liet zien dat het in de praktijk niet gemakkelijk is om voorkennis te verzamelen en om te zetten in prior verdelingen. Het hoofdstuk eindigt daarom met een aantal richtlijnen voor onderzoekers om mee te werk te gaan. Nieuwe toegepaste studies waarin voorkennis wordt verzameld en omgezet in

prior verdelingen zullen meer licht werpen op dit proces en onderzoekers verder helpen in het gebruiken van voorkennis.

Hoofdstuk 4, waarin een procedure werd ontwikkeld om voorkennis bij experts te verzamelen, toonde aan dat maatwerk hierin van groot belang is. De procedure moet aangepast worden op de inhoudelijke en statistische kennis van de expertgroep in kwestie. Veen et al. (2017) heeft voortgeborduurd op het werk van Zondervan-Zwijnenburg et al. (2017b) en ontwikkelde een digitale variant van de sticker-methode waarmee de experts hun voorkennis weergaven. Deze methode kan de validiteit en betrouwbaarheid van de methode verder verhogen doordat de experts onmiddellijk terugzien hoe hun gestickerde verdeling wordt omgezet in een statistische verdeling. De methode van Veen et al. (2017) kan opgenomen worden in de procedure zoals verder voorgesteld in Zondervan-Zwijnenburg et al. (2017b). Op welke manier de informatie verkregen wordt (bijv. online, face-to-face, of face-to-face met behulp van computers) hangt wederom volledig af van de experts waarbij de informatie verkregen wordt. Sommige experts zullen betere informatie geven wanneer zij in een groep zijn en vragen kunnen stellen, terwijl andere experts mogelijk beter functioneren buiten een vergadering om.

In Deel II van de dissertatie stond het testen van replicatie met de 'prior predictieve p-waarde' centraal. Deze methode neemt steekproefvariantie mee en is op een breed scala van statistische modellen toepasbaar. Het testen van beweringen uit een originele studie die verder gaan dan een effectgrootte was met geen van de bestaande methoden om replicatie te evalueren mogelijk. De uitbreiding naar statistische modellen naast de ANOVA in Hoofdstuk 6 was een belangrijke stap in de huidige replicatie literatuur. Twee R-packages en een online R-Shiny omgeving zijn gemaakt om de methode goed bruikbaar te maken voor onderzoekers. Vervolgstappen kunnen gemaakt worden omtrent de verwerking van missende data binnen de replicerende studie en het berekenen van statistische power voor modellen anders dan de ANOVA.

In Deel III tonen we aan dat cohort-studies goed gecombineerd kunnen worden door ondersteuning voor hypotheses samen te vatten over studies heen. Deze methode kan ook op andere modellen en onderzoeksvragen toegepast worden om robuuste ondersteuning van hypotheses te evalueren.

Acknowledgements

En nu wij zijn omringd
door zoveel helden die ons voor zijn gegaan,
nu geven wij niet op.
Het zijn de schouders waarop wij mogen staan.

Tim Hughes, Kees Kraayenoord, Matthijn Buwalda

First, Rens and Herbert, without you, this dissertation would not exist. Rens, thanks for all the opportunities that you provided, the experience that I could gain as a student-assistant, and your coaching along the way. You have taught me to see new opportunities in every set-back, and your optimism has always been encouraging. It was great to be part of your group. Herbert, thanks for your trust in me and this project. Thank you for your questions, your thinking, your explanations, and your feedback. I have been very lucky with two supervisors who were so supportive and committed: the emoticons in your e-mails were much appreciated.

Kimberley, it was great to have you as a roommate for so long. The food cultivation experiments on your desk were extremely interesting. Thank you for all the good conversations, and your sweet and funny personality. Sanne, Duco and Marthe, next to Kimberley you were also in Rens' group and you have been very important to me. Thank you for your never-ending kindness and humor. I am grateful for all the fun, and efficient Monday morning meetings that we had together (and the cookies that were automatically passed in my direction). Sanne, you are the best backup-paranimf!

I'd like to thank everyone in the M&S department for making it such a warm place to be. Every event with cookies, cake, and colleagues has been great! Joop Hox, thank you for giving me your "Bayesian" and "statistics" refrigerator magnets when I was still a DaSCA student: it was a prophetic gift. Kevin, Flip, Chantal, Marianne and Els: you make everything happen.

Not only the UU M&S department is full of great people. I'm also really happy that I got to know my M&S research master classmates and two classes of DaSCA students, including my paranimf Rianne, who manages to combine fun, family and great ambitions in an inspiring manner. And then there's the Consortium Individual Development including Lotte, Sanne, Yasin, Jiska, Michelle, Stefania, Alex, Stefanie, Nathalie, Andrik, and my fantastic master's thesis supervisor: Margot. It has been great to collaborate and/or drink hot chocolate with you. Sabine, we 'only' collaborated in Part III of this dissertation, but what a project and what a collaboration that was! Not only did your questions and checks keep me sharp, on a personal level it was also wonderful to work with you.

Papa en mama, bedankt voor het vertrouwen en alle ruimte die jullie mij en ons hebben gegeven. Bij jullie ben ik altijd thuis. Dankzij jullie heb ik een geweldige basis om me mee te kunnen ontwikkelen. Ik hoop dat ik hetzelfde door kan geven. Vrienden en (schoon)familie, bedankt voor jullie interesse, meeleven en gezellige afleiding op z'n tijd.

Finally, Rick, you have always been there and supported me like a super giant. You are my favorite husband, and I am glad to be in your team. Elianne and Jaïr, you definitely did not help in writing this dissertation, but you did help me develop my resilience by making me live in the moment. Who cares about publications when we need to sing, read, chase and tickle?

Woerden, July 2019
Mariëlle Zondervan-Zwijnenburg

About the Author

Mariëlle Zondervan-Zwijnenburg was born to Jacob and Marian Zwijnenburg at December 24, 1989. She was raised with her four siblings at the countryside near Polsbroekerdam. Here, Mariëlle explored futures in playing the clarinet (not much appreciated by the family) and fierljeppen (not very successful).

After her secondary school, Mariëlle finished a bachelor in Pedagogical Sciences. In her third year, however, she noticed that she liked research better than clinical practice. Hence, she enrolled in the Utrecht University research master: Development and Socialization in Childhood and Adolescence (DaSCA).

During that time, she became a student-assistant for Rens van de Schoot at the department of Methods and Statistics. He showed how statisticians can do useful work, which made that Mariëlle enrolled in the research master of Methodology and Statistics of Social and Behavioral Sciences (M&S) as well.

In May 2014, Mariëlle graduated cum laude for both research masters. Subsequently, she started her PhD within the Consortium Individual Development (CID) of which this dissertation is the main result. In the meantime, Mariëlle also gave birth to two children: Elianne in April 2016, and Jaïr in December 2017.

Mariëlle is passionate about working on the bridge between statistics and social sciences. Her ambition is to enable social science researchers to use the best methods to optimally analyze their data and answer their research questions. Teaching, presenting, writing papers and easy-to-use R-packages are important manners to accomplish this goal.

International Peer-Reviewed Publications

- **Zondervan-Zwijnenburg, M.A.J.***, Veldkamp, S.A.M.*, Neumann, A., Barzeva, S.A., Nelemans, S.A., Van Beijsterveldt, C.E.M. Branje, S., Meeus, W.H.J., Hillegers, M.H.J., Tiemeier, H., Hoijtink, H.J.A., Oldehinkel, A.J., & Boomsma, D.I. (2019). The impact of parental age on child behavior problems: Updating evidence from multiple cohorts. *Child Development*. doi: 10.1111/cdev.13267
  * These authors contributed equally.
- **Zondervan-Zwijnenburg, M.A.J.**, Depaoli, S., Peeters, M., & Van de Schoot, R. (2019). Pushing the Limits: The performance of ML and Bayesian estimation with small and unbalanced samples in a latent growth model. *Methodology, 15*, 31-43. doi: 10.1027/1614-2241/a000162

- Veen, D., Stoel, D., **Zondervan-Zwijnenburg, M.A.J.**, & Van de Schoot, R. (2017). Proposal for a five-step method to elicit expert judgement. *Frontiers in Psychology, 8*, 2110. doi: 10.3389/fpsyg.2017.02110
- **Zondervan-Zwijnenburg, M.A.J.**, Peeters, M., Depaoli, S., & Van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development, 14*(4), 305-320. doi: 10.1080/15427609.2017.1370966
- **Zondervan-Zwijnenburg, M.A.J.**, Van de Schoot-Hubeek, W., Lek, K., Hoijtink, H., & Van de Schoot, R. (2017). An expert judgment elicitation procedure for correlations. *Frontiers in Psychology, 8*, 90. doi: 10.3389/fpsyg.2017.00090
- Van de Schoot, R., Winter, S.D., Ryan, O., **Zondervan-Zwijnenburg, M.A.J.** & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217-239. doi: 10.1037/met0000100
- Van Dijk, R., Deković, M., Bunte, T.L., Schoemaker, K., **Zondervan-Zwijnenburg, M.A.J.**, Espy, K. A., Matthys, W. (2017). Mother-child interactions and externalizing behavior problems in preschoolers over time: Inhibitory control as a mediator. *Journal of Abnormal Child Psychology, 45(8)*, 1503-1517. doi: 10.1007/s10802-016-0258-1
- Van de Schoot, R., Broere, J., Perryck, K., **Zondervan-Zwijnenburg, M.A.J.**, & Van Loey, N.E.E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology, 6*(25216), 1-13. doi: 10.3402/ejpt.v6.25216
- Van de Schoot, R., Schmidt, P., Beuckelaer, A.D., Lek, K., & **Zondervan-Zwijnenburg, M.A.J.** (2015). Editorial: Measurement Invariance. *Frontiers in Psychology, 6*, 1064. doi: 10.3389/fpsyg.2015.01064
- Peeters, M., **Zondervan-Zwijnenburg, M.A.J.**, Vink, G. & van de Schoot, R. (2015). How to handle missing data: A comparison of different approaches. *European Journal of Developmental Psychology, 12*(4), 377-394. doi: 10.1080/17405629.2015.1049526

Book Chapters

- **Zondervan-Zwijnenburg, M.A.J.**, & Rijshouwer, C.D.N. (2020). Testing Replication with Small Samples: Applications to ANOVA. In: R. van de Schoot, M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners. Routledge*

Other Publications

- **Zondervan-Zwijnenburg, M.A.J.** (2019). How to Test Replication for Structural Equation Models. *PsyArXiv*. doi: 10.31234/osf.io/uvh5s

- **Zondervan-Zwijnenburg, M.A.J.**, Van de Schoot, R., & Johnson, A. R. (2017). Computing complexity for the Bayes Factor in inequality constrained hypotheses. doi: 10.17605/osf.io/5YT3J
- **Zondervan-Zwijnenburg, M.A.J.** (2015). Ontwikkeling van jonge cannabisgebruikers vergeleken met leeftijdsgenoten: Een Bayesiaans avontuur. *STAtOR, 15*(2), 4-9. Persistent-Identifier

CRAN R-packages

- **Zondervan-Zwijnenburg, M.A.J.** (2019). Replication: Test Replications by Means of the Prior Predictive p-Value. R-package version 0.1.0. `https://CRAN.R-project.org/package=Replication`
- **Zondervan-Zwijnenburg, M.A.J.** (2018). ANOVAreplication: Test ANOVA Replication by Means of the Prior Predictive p-Value. R-package version 1.1.3. `https://CRAN.R-project.org/package=ANOVAreplication`
- **Zondervan-Zwijnenburg, M.A.J.** (2017). complexity: Calculate the Proportion of Permutations in Line with an Informative Hypothesis. R package version 1.1.1. `https://CRAN.R-project.org/package=complexity`

Awards

- Peter G. Swanborn Research Master's Thesis Award 2013-2014. Award to promote high quality social science research by students: Research that is characterized by attention for theoretical as well as empirical components, by creativity and applicability, and by high quality. €1,000.- `http://www.uu.nl/organisatie/faculteit-sociale-wetenschappen/peter-g-swanbornprijs`