Bayesian Hierarchical Latent Mixture model for Meta-analysis

BHLM R package introducing Bayesian Latent Mixture Model as a method for meta-analysis: Issues facing modern research and meta-analysis and how Bayesian methods might help solve them.

Hugh Benjamin Zachariae

Aarhus University

May 30th, 2018

Abstract

Scientific research is currently under much pressure from the scientific community as recent studies have indicated that common methods do not give replicable findings. The criticism has recently accelerated with the “replication crisis” of Psychology and the increasing frustration with publication bias. With its increase in popularity, meta-analysis has recently stood as a plausible solution to this crisis of scientific method. However, there are still issues that meta-analysis seem to struggle to solve. With the increase in computational power, Bayesian ideas pose radically different approaches that could solve many of the crisis issues elegantly, but Bayesian methods are still underutilised in meta-analysis. This paper introduces and overviews the BHLM package to R, which utilises hierarchical latent mixture analysis as an approach to meta-analysis. Using a prototype build of the package, I re-analyse two recent meta-analysis studies. The analyses produce negligible effects, similar to the original papers, which supports the validity of the package and the method. However, the method allows for more elaborate analysis based on further developed and theorised priors. In the end, future implementations to the package are discussed. These include, Bayesian bias mitigation and assessment and the package as a stepping tool to introduce Bayesian methods to more researchers. Such introductions, ideally, could lead towards a solution to the current scientific crisis.

Keywords: Bayesian inference, Meta-analysis, Hierarchical latent mixture model, Reproducibility crisis, Publication bias, R.

Introduction

In recent time, psychological research has been subject to a lot of scepticism. Most recently, the Open Science Collaboration (2015) published its estimate on the reproducibility of psychological science and the results caused a big uproar in the scientific society. Most often, the reliance on *p* values, especially in publishing, has been pointed out as the largest contributor to the current crisis (Wagenmakers, 2017). The effects of publication bias have been shown by Masicampo and Lalande (2012) as a peculiar prevalence of p-values just below the threshold (.05) in three prominent psychology journals. A strong claim was made by John Ioannidis (2015) that "most published research findings can be proven false". By analysing the post-study probability that a research finding is true, Ioannidis argued that research fields that had the following properties; being mainly informative (i.e. hypothesis-generating), based on low statistical power, characterized by high analytical flexibility (thus fuelling biased results), and being highly popular, correlated with a very low probability of revealing actual true findings. Ioannidis especially underlined the large effect of bias on the probability of the findings being true. These issues support the fact that scientific method is fallible, unlike what many want to believe and that we need to be constantly critical of current research practices and results. However, while these issues may have their downsides, it can be said that criticism and crisis fuels development. Such developments, for example, can be seen in the surge in popularity of meta-analysis and systematic reviews.

Meta-analysis is defined in the social sciences as "The statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Haidich 2010, Glass 1976). Meta-analysis was developed to overcome the starting inadequacies of traditional summarizing methods when approaching an ever-building volume of research (Glass 2015). When we try to study processes that are too complex to yield simple answers, a hundred studies will often return a dozen varying results. When "ground-breaking" effects cannot be reproduced, some think that the answer may lie in compiling the data. Meta-analysis includes the study of the variance of these results to better answer the question "how large the particular effect is". Since the seventies, meta-analysis has grown into a large influence, especially in medicine and with the current issues afflicting research and many see it as a plausible solution.

However, criticisms are still levelled at meta-analysis. Most prominent is the issue of missing variables – both truly missing statistics like standard deviations or effects believed to be missing,

caused by publication bias. More recently, Bayesian methods have gained more traction in statistical research, mostly due to increasing computational power, but also as a counter-movement to the issues afflicting frequentist methods. Bayesian methods could provide solutions to some of its issues.

Unknown quantities

Missing or unreported statistics is an issue that will have troubled any researcher who has run a meta-analysis, especially missing standard deviations. Common methods to handle missing values are to contact the researcher for missing data, using reported statistics to recalculate the missing statistic, or imputing a single value – e.g. setting missing standard deviations as the baseline standard deviation (Carpenter & Kenward, 2013). Especially with very large or old fields, it can be a trying endeavour to personally dig up the original statistic and often success here can be very limited. Recalculating the missing statistic can also be impossible, as it is limited by the availability of often several other analytical results being available. Imputing a single value under assumptions may distort future analyses by ignoring the uncertainty around it (Grant, 2018).

Bayesian methods use probability distributions to describe unknowns. With probability distributions the uncertainty can be modelled according to prior knowledge or a specific scientific question. For example, you can model a specific variance, theorised according to prior knowledge, or you can input a vague or uninformative prior which would respect the fact that we might not have any guaranteed prior knowledge (e.g. a uniform prior or wide Gaussian distribution). Bayes theorem allows us to update our belief and approximate a posterior hypothesis as we introduce more information. Emulating the uses of *p* values, a "Bayes factor" can be calculated to estimate the increase in probability on a point of interest and answer whether we need to change (or update) our belief according to the observed data.

Publication bias

As mentioned, a major challenge in meta-analysis is the so-called file-drawer problem or the risk of publication bias (Borenstein et al, 2011). There is considerable evidence suggesting that studies are more likely to be published when they report large, statistically significant effect sizes, which may skew the available evidence in the direction of overestimated effects. The issue is most often associated with the arbitrary threshold for *p* values in null-hypothesis testing that seem dictate whether results are published or not. As mentioned, this issue is only further underscored in

psychology during recent years by the so-called "reproducibility crisis" (Open Science Collaboration, 2015). The undue emphasis on statistical significance could be one explanation for this "crisis" as *p*-values fail to convey information about likelihood of replication. Furthermore, the threshold can push researchers to feel pressured towards attaining significant results, increasing the probability of other research biases (e.g., repeated-peeks-bias[1]). One bonus to gain from Bayesian methods is that no corrections are needed for such optional stopping, as there are no corrections needed for continuing or stopping the data collection (Wagenmakers, 2007; Edwards et al. 1963), especially if we have well defined priors – a blessing and a curse which will be discussed later in the paper.

While methods have been developed to deal with bias in meta-analysis, often they are only used as post hoc (after the fact) assessments of sensitivity and possible bias (Grant, 2018; Lau et al, 2006). The most common method to assess bias is the funnel plot. The method plots effects on the horizontal axis and some measure of precision on the vertical axis. If all studies come from a single underlying population, it should look like a funnel, with the effect homing in on the true as the sample size increases (Light & Pillemer, 1984). Asymmetry in the plot is seen as a diagnosis of missing studies, or publication bias. However, it is still unclear whether a funnel plot is a true diagnosis of publication bias, as the asymmetry might not be developed from unreported effects, but from an underlying difference between the studies with small and large sample sizes (Wagenmakers, 2007). Methods have also been developed to mitigate the effect of bias, e.g., the trim and fill method by Duval and Tweedie (2000). This method imputes the hypothesized missing values from the funnel plot. However, as discussed above, imputing single values could distort the analysis by ignoring uncertainty of these values.

Bayesian method is a probability-based statistical analysis. This means that any result from a Bayesian analysis is affected by our prior belief. This fact allows us to try modelling the bias – e.g. with censoring functions proposed by Guan and Vandekerckhove, 2016. Alexander Etz and Vandekerckhove (2016) has developed a Bayesian method of mitigating the effect of publication bias for fixed effects meta-analysis and extended to random-effects meta-analysis via a hierarchical

[1] Capitalising on random fluctuations in data collection. Monitoring the data until the effect or level of significance appears to show (Masicampo & Lalande, 2012, Wagenmakers, 2007).

model. Their bias correction method models the publication filtering process which causes publication bias and modifies the sampling distribution (*g* in our model) to describe such assumptions more accurately. Etz (2017, May 21) demonstrates the method in a recent lab presentation and uses the four censoring process assumptions, which describe no bias, extreme bias, constant bias, or exponential bias[2]. He implements these assumptions in a Bayesian model which infers the true effect under such bias assumptions. His results show that under such bias assumptions, the difference between large and small effects shrink towards the overall mean. The Bayesian bias mitigation method was shown to produce results closer to the true effect size in simulation experiments (Guan & Vandekerckhove, 2016). The method was also shown to have a strong effect on driving the evidence more towards the null hypothesis in meta-analysis studies where a low number of non-significant studies could implicate a publication bias.

In all, Bayesian methods seem to have strong inherent abilities and developed models towards handling publication bias. However, Bayesian methods are still very underutilised in meta-analysis (Grant, 2018). Therefore, this paper sets out to introduce a Bayesian method of meta-analysis, utilising hierarchical modelling and latent mixture analysis. The process of the method is implemented in a package, BHLM, for the statistical program R and is meant to automate the process while not undermining the importance of very well considered priors. Before describing the package, I will give a brief introduction to Bayesian inference, needed to understand the underlining theory and ideas behind the package.

Introduction to Bayesian inference

The basis of Bayesian thinking is that idea that we cannot know exactly what an event will result in. No matter how many tests or data collections we do, we can never repeat that process the infinite times needed see all possible outcomes. Instead we can look at potential outcomes in terms of probabilities and we can update these probabilities, our belief, according to observed evidence.

[2] No bias meaning that all results are published, extreme bias meaning that no non-significant results are published, constant bias, meaning that non-significant results are published at a certain constant rate, and exponential bias, meaning that non-significant studies are published more, when closer to .05 *p* (Guan & Vandekerckhove, 2016).

For example, myth states that the invention of Bayes theorem, came when Thomas Bayes[3] sat with his back to a table and asked his assistant to throw a ball onto a table (McGrayne, 2011). He would then ask his assistant to throw another ball onto the table a tell him where new ball landed in relation to the first. Bayes would then compile the information after many balls thrown at the table to update the probability of where the original ball was located. The experiment explains well the idea that we can never know the world perfectly, and were the ball is located, but we can only hope to update our belief according to observed data. This idea is the basic feature behind Bayesian results and interpretation and, personally, this is what draws me to Bayesian statistics. The frequentist approach to inference, simplified, requires all probabilities to be tied to observed events, meaning that if we observe enough data, a pattern will emerge that describes the effect. The distribution of a data sample is called the sampling distribution[4]. This means that uncertainty is premised on imaginary resampling of the data. However, where frequentist ideas see this distribution as deterministic, non-random noise (as it comes from the measurements), Bayesian inference models this randomness with probability distributions. In theory, if we had all information available to us, we should in theory be able to predict all things. Bayesian ideas tackle our lack of ability to gather all information, by letting this randomness describe this lack of knowledge (McElreath, 2016, p. 11) and this is done via the posterior probability distribution, updated from the prior probability distribution via Bayes' theorem.

Put simply, a Bayesian model has prior information on the probability of specific data outcomes occurring. As the model is then fed data, it updates the probabilities according to the observed outcome. Mathematically, this is done via Bayes theorem.

$$p(\theta|X,\alpha) = \frac{p(X|\theta) * p(\theta|\alpha)}{p(X|\alpha)}$$

[3] The idea was published by Richard Price and further developed by Pierre-Simon Laplace.
[4] The sampling distribution is equal to the distribution of *g* in our model (e.g. where we have our observed data) and should not be confused with the posterior or prior distribution.

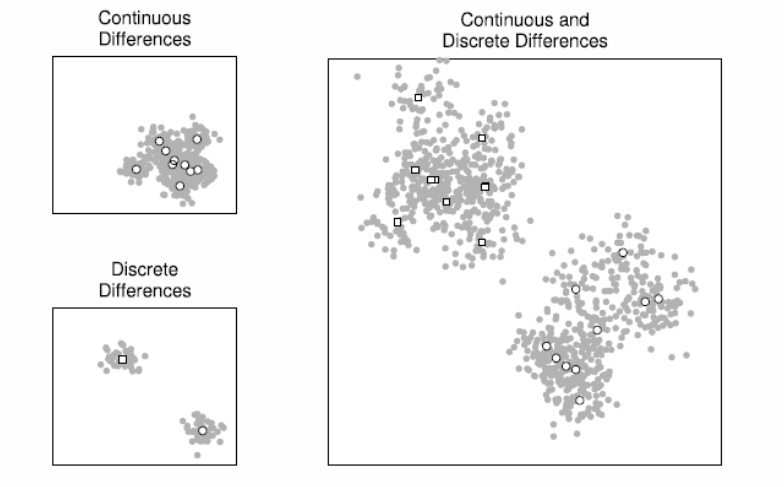
This equation states that the posterior distribution ($p(\theta|X,\alpha)$) is equal to likelihood ($p(X|\theta)$) times the prior distribution ($p(\theta|\alpha)$), over evidence. Practically, modern Bayesian methods uses Markov Chain Monte Carlo estimation to approximate the posterior distribution via the Bayes theorem.

Dependency on the prior distribution

As with all statistical methods there are issues and critiques to Bayesian statistics. McElreath (2016, p. 20) explains; "The way Bayesian models learn from evidence is arguably optimal in the small world. When their assumptions approximate reality, they also perform well in the large world.". One strength of Bayesian inference is that sample size is irrelevant, in terms of getting valid interpretation of the results from the analysis. Even with few data points, you can update a prior distribution and get a posterior distribution. Even when these distributions prove to be extremely similar, this only means that we have not observed enough evidence to change our belief beyond the prior we set. From that point, you can Bayesian methods allow you to continue collecting data and update your knowledge with more data which ideally should increase our precision in predicting the result. However, this exposes weakness of low sample size in the dependency on the prior – which, as mentioned, is both a blessing and a curse. If you have a small sample size and a prior which is not true to the real world, you will most often get misleading results, in terms of describing the real world. However, this does not mean that such results are not beneficial. Given that we have made a commitment to a model, we *can* be sure of the estimate it produces. In this, models can be designed to answer specific questions. However, this makes it very important in Bayesian modelling to critique and supervise the model. This especially means, defining priors in a way that is relevant to specific questions and grounded in a reasonable knowledge of the real world and its probabilities. However, before priors are defined we need to

Hierarchical Latent Mixture model

Figure 1: Three different modelling assumptions about individual differences (Full difference and no difference have been removed).



Bartlema et al. 2014.

The model used in this package, is based on two concepts: Hierarchical modelling and mixture modelling. In figure 1, is shown five different assumptions about individual differences. The circles and squares indicate the true point and the grey region surrounding it shows the inferences that can be made from data collected. The top-left panel shows the assumption of continuous individual differences between participants that vary around a central tendency. Hierarchical modelling uses multi-levelled structure to accommodate this difference. This means that any individual mean is conditional on a group level mean to describe the central tendency. The bottom-left panel shows the assumption that there are individual differences, but that they are due to a more particular difference between latent groups, shown by the square and circle symbols for each group. The right panel shows the combination of these two assumptions, which then would be best described by the combination of the two modelling approaches.

Bartlema (2014) argued and demonstrated that hierarchical mixture models reveal clear individual differences and thus provides a more accurate and complete understanding than can be gleaned from an analysis based on aggregate data. Meta-analysis studies use aggregated means across studies to assure independence between outcomes, however, hierarchical mixture modelling should provide a more accurate picture of individual differences between studies and the latent variables modelled therefrom.

BHLM package prototype build

The BHLM package has been developed to fit well with intervention meta-analysis studies. Especially, towards meta-analyses that include studies with many outcomes based on many different methods and researchers. An example of this setup can be seen in table 5 and 8. Using hierarchical modelling and latent mixture analysis with Bayesian inference (see figure 2), allows researchers to infer towards latent outcome variables while accounting for individual study difference.

The BHLM package was built for and using R and is currently available in its current prototype build (under the current alias "ZachariaeBHLM"), downloaded from via Github.

```
install.packages("devtools")
devtools::install_github("https://github.com/arcuo/bachelors_meta")
```

The current build of the package includes three functions, a main "bhlm" function, two plotting functions, one to review trace plots and one to produce posterior/prior distribution plots according to the Savage-Dickey method (Wagenmakers et al, 2010), and one function to produce maximum a posteriori (value with the highest probability) with different density estimate methods.

BHLM function

```
bhlm(dataframe, grouping_factor_cols, estimate_col, outcome_options_col,
     outcome_options,
     outcome_priors, lambda_prior, theta_prior,
     field_theta_precision = NULL,
     identifier_col = NULL,
     bayes_method = "jags",
     jags_init = NULL, jags_chains = 3, jags_iter = 2000,
     jags_burnin = floor(jags_iter/2),
     jags_thin = max(1,floor(jags_chains * (jags_iter - jags_burnin)/1000)),
     jags_DIC = TRUE,
     save_model = NULL)
```

The BHLM-function is the main function of the package. This package uses helper functions (not exported for use, at least in the current build) to, first, prepare the given data frame by removing variables that are not used while calculating and creating data vectors used in the JAGS method. Second, the function converts outcome priors (`outcome_priors`). Third, the function creates the

model file for JAGS. This text-file is either temporary (using the `tempdir` and `tempfile` functions which comes along with R) or saved as to the path defined with the `save_model` parameter5. The option to save the implemented model allows total transparency when reporting the model, which should be an encouraged habit. Therefore, this feature is by default set to save the file in the working directory. Next, the function samples the data and tries to approximate a posterior distribution for each latent mixture variable. For R, there are currently two samplers available. First, there's JAGS (Just Another Gibbs Sampler), which uses a variant of the Hastings-Metropolis algorithm. We connect to the JAGS engine through the package R2Jags (Su & Yajima, 2015). The relative simplicity of the model used in the BHLM package, makes JAGS very efficient in approximating a posterior distribution. Secondly, there's Stan, which uses the Hamiltonian Monte Carlo (HMC) method of sampling. HMC is a more recently developed method and which has been shown to be more efficient as the number of parameters and the complexity of the model increases. Stan can be accessed with R through the package RStan (Stan Development Team, 2018). Currently, the package is only implemented with R2Jags and does not run HMC. This will be a future introduction to the package. Currently, function creates samples using the `jags` function from the package R2jags JAGS results, used data, loop starting bounds, outcome prior distribution specification, and estimate variable name is saved in a custom `bhlm_object`. This object allows easy and automated use of the plotting functions of the package.

While the number of arguments might seem daunting, most of them have default specifications to allow easy use, while allowing more experienced users to customise the model, the JAGS-procedure, and saving the model. The necessary arguments are the data frame that the function will use along with naming specifications for grouping factor columns (study number and outcome number in the study), the estimate column (e.g. the column with Hedge's *g*), and the latent variable column. With `outcome_options` you can choose the specific latent variables to be used from the latent variable column. For custom plots, it's possible to define as column in the data frame as

[5] In the future, this method will be deprecated to change the use of temporary files to instead just use strings, ensuring that the function works on all operating systems without having to access the filesystem.

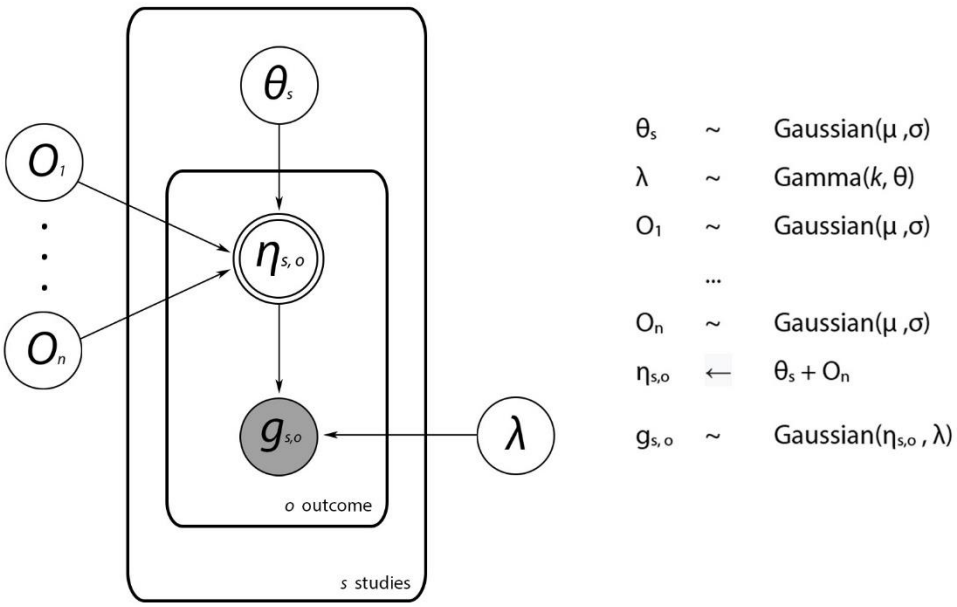
the identification column (e.g. a column with study author name(s), etc.). The data defined with these arguments are returned by the function as the `used_data` variable.

Because of the large effect of the prior distribution in Bayesian inference, especially when we have only few outcome data points, priors are given no default definition to ensure that priors are considered in terms of prior knowledge and research question when running the analysis. The function allows the user to define one prior distribution for all outcomes, but warns that the user should, optimally, define each prior manually. Optionally, it's possible to add another hierarchical level to the model with a "field" mean, where you select. Uses of this option will be discussed later in the present paper.

BHLM model specification

The model is designed with a hierarchical structure to consider the difference between studies and a latent mixture variable structure, so that we can infer the latent variable distributions. In the inner loop (over outcomes), the observed effect variable *g* can be any effect variable type, theoretically, however is denoted as Hedge's *g* for the examples which use effect size. The sampling distribution of *g* is constrained by a mean η and a precision (inverse variance) λ. The mean η is constrained both by the hierarchical and the latent mixture structure. The hierarchical structure; study mean θ ensures that individual study differences are considered while giving a central tendency distribution in the prior distribution. The θ prior distribution is most sensible with a mean of 0 to ensure assumption that η is controlled by the latent variables. By this assumption, the study does not produce the effect, but the latent group variables are generating the effect. However, the prior for θ is open to be determined according to any knowledge of bias. However, the structure of η (θ + *O*) makes it so that a higher mean of θ will lower the effect of latent variables (which could be meaningful when considering study effect bias). The current structure of the implementation allows only one prior distribution for all studies. Future implementation of separate study priors is discussed later the present paper. The latent mixture structure; makes it so that the mean of the effect distribution is also constrained by the latent variable distributions. This allows us to infer the posterior distributions of each of the latent variables. Currently, the sampling distribution variance is defined as a single prior distribution. Further inclusion of the variance parameter into the hierarchical structure with observed standard error is also discussed later in the present paper.

Figure 2: Hierarchical Latent Mixture Model.



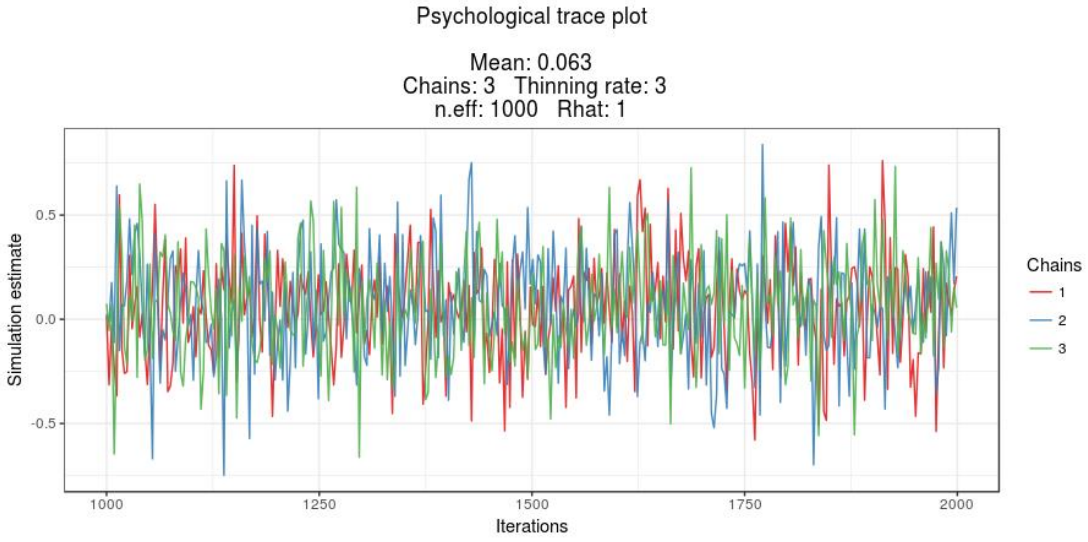
θ = study specific effect, η = Mean of effect difference, λ = conjugate prior for effect sampling distribution variance (gaussian), O = latent variable and *g* = observed effect size sampling distribution. Note: In JAGS, variance is defined with precision (the lower the value

The implementation of the model is seen in figure 5 and 6, appendix for each study example.

Trace plots function

```
bhlm.traceplot(bhlm_object, outcome_options = NULL, return_plots = FALSE)
```

Figure 3: Example trace plot from latent variable "Psychological" from EWI study.



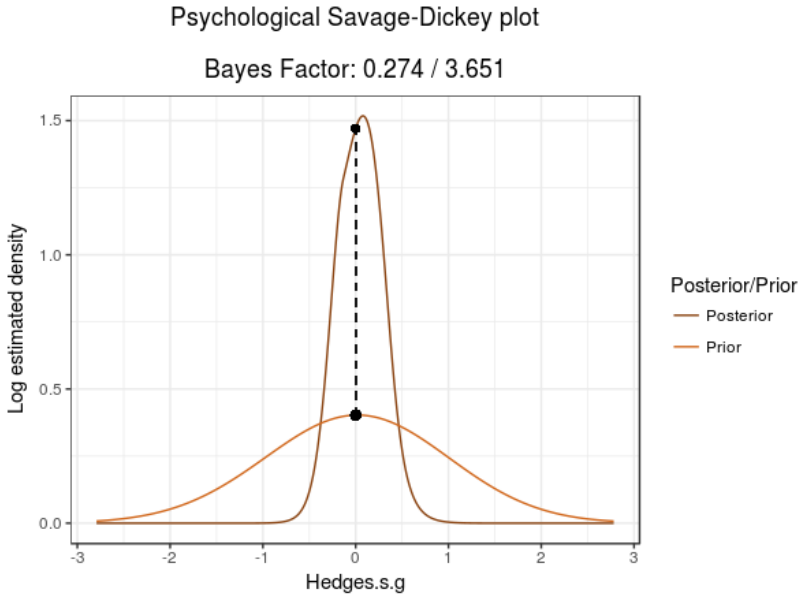
The trace plots made by `bhlm.traceplots` show the path of the sampling chains (not including the burned samples made during the adaption period[6]) and are a good tool to examine the health of the sampling chains produced by the sampler. The function allows you to plot specific outcome variables ("theta", "lambda", and/or any number of the latent variables) with the `outcome_options` argument. Defaults to the latent variables. If the user wants to return the plot objects for further customisation, he she can set `return_plots` to true. Figure 3 presents an example of a trace plot produced by the function (from the EWI example study). The plots allow the user to review four indicators of a healthy chain: Stationarity, good mixing, Gelman-Rubin convergence diagnostic (Gelman & Rubin, 1992), Rhat, and effective samples (n.eff) (McElreath, 2016, p. 253). Stationarity means that the sampling stays within the posterior distribution with a central. Visually the simulations should bounce around the posterior mean. Unhealthy stationarity could be that the mean changes over time (e.g. an upward trend or a sudden changes). Good mixing means that the samples are not highly correlated with the sample before it. This is shown visually by the rapid zig-zagging motion of the chain without the path getting mired at any sampling value. If there are iterations with low variety, there could be issues with mixing in the sampling process. The Gelman-Rubin convergence diagnostic indicates whether the chains have converged. A "Rhat" above 1.00 usually indicate that the chain has not converged. The number of effective samples indicate the effectiveness of the chain. If "n.eff" is much below the number of unburned iterations, this means that the chain is not very efficient. McElreath (2016, p. 257) suggest that you should not rely too much on the two latter statistics as they might give wrong advice on the convergence (both positively and negatively). The more complicated the model, the less we can trust these diagnostics. However, it is good should be pushed that it is scientific practice to review and report these statistics.

[6] Number of burns is set by the argument jags_burnin. This means that the first samples are for the model to adapt and become more efficient at sampling the posterior distribution.

Savage-Dickey plots function

```
bhlm.SDplots(bhlm_object, null_hypothesis,
             outcome_priors_data = NULL,outcome_options = NULL, return_plots = FALSE,
             density_estimation = "logspline", cum_prob = NULL,  iter = 10000)
```

Figure 4: Example Savage-Dickey plot from latent variable "Psychological" from EWI study.



The Savage-Dickey method is a framework for Bayesian hypothesis testing which is easily carried out and easy to interpret (Wagenmakers et al, 2010). The method consists of calculating the Savage-Dickey density ratio (Dickey & Lientz, 1970), or Bayes factor, between prior and posterior distributions at the point of interest. The Bayes factor describes the change from prior odds to posterior odds and is most commonly used as the weight of evidence towards the posterior. If you interpret prior and posterior probability distributions as knowledge, this weight determines whether we need consider changing our belief from a prior belief to the posterior belief. For the bhlm-package we mostly want to investigate the effect produced by the latent variables. Therefore, the `bhlm.SDplots` defaults to testing the posterior distributions of each of the latent variables with producing a Savage-Dickey plot for each. For distribution smoothing and estimation, the function allows the user to choose between three types of density estimation. First (Default) estimation with the "polspline" package (Kooperberg, 2015) to estimate the log-density using logsplines. This method is shown to have both great accuracy when estimating the density (Deng & Wickam, 2011). Second, estimation with the density function coming with R base (R Core Team,

2018) with default settings. Third, the function plots with the ggplot2 package function `geom_density` (Wickam, 2009). This method is not fully implemented yet and throws a warning. In the future, it could be feasible to incorporate the more choices of density estimation tools. This could include methods by the "ash" package (Scott, Gebhardt & Kaluzny; 2015) or the "KernSmooth" package (Wand, 2015), both of which were highly praised by Dern and Hadley. If needed, the function can return both the plots (ggplot2), the data used for the plot, and the density estimation objects. The function automatically samples the prior distribution for each outcome, according to the settings saved in the `bhlm_object`. Because JAGS uses precision instead of standard deviation, the function re-calculates the standard deviation from the precision given in the settings as seen below.

```
prior <- dnorm(mean, precision) # defined in argument outcome_priors of bhlm
prior.sampling <- rnorm(n, mean, 1/(precision)^2)
```

Maximum a posteriori function

```
bhlm.MAP(bhlm_object, outcome_options = NULL, density_estimate = "logspline")
```

The maximum a posteriori estimate (MAP) is the mode, most probable sample or top of the density distribution, is most often used to get a grasp of the actual inferred effect for Bayesian inference. One critique is, as with most Bayesian methods, that posterior distributions might be very influenced by the prior, meaning that the MAP estimate will not give us a result that is true to the real world if the prior is not correct. In this case, it might not be a useful estimate for research. Another critique is that the MAP estimate might not be feasible in cases where the posterior is multi-modal. However, for this model, this should not be the case, as it is not a mixed-model and should, most often, not produce multi-modal posterior distributions. The function allows the user to choose between MAP estimated from the base density function of R density estimate, or the log density estimate of "logspline". Choice of method defaults to log estimation, but users should consult the Savage-Dickey plots of each method for a visual representation. As with the "SDplot" function, future methods should be added here.

Method

To give an example and proof of concept to the usage and results of the package, I will overview two meta-analysis studies in psychology. Both studies showed mostly negative and insignificant

results. These studies should demonstrate well how the package can be used for meta-analysis studies. The studies' data and results are overview below.

Study 1 – Zachariae & O'Toole, 2015, The effect of expressive writing intervention on psychological and physical health outcomes in cancer patients – a systematic review and meta-analysis

The aim of the study was to evaluate the effectiveness of expressive writing intervention (EWI) on psychological and physical health in cancer patients and survivors. The meta-analysis included studies (a) with a study population of adult cancer patients or survivors, (b) that used EWI methods according to the Pennebaker paradigm (Pennebaker & Beall, 1986;95), (c) randomized participants to EWI and one or more control conditions, (d) presented data for both EWI and control groups on the related outcome types (including Quality of Life, QoL), and (e) reported results as pre-post means, standard deviation/error, change scores, effect size (e.g. Cohen's *d*), or other relevant statistics, for all groups. The head of the data (the first 5-10 rows) is shown in table 5 (appendix) (columns that were irrelevant to the analysis conducted in this paper are omitted).

The study reported several meta-analytic results for several outcomes. The outcomes that will be compared in the present paper are *Psychological health combined* (Hedge's *g*: 0.04, Cl: -0.06 to 0.14, *p* (two-tailed): 0.419), *Physical health combined* (Hedge's *g*: 0.08, Cl: -0.05 to 0.20, *p* (two-tailed): 0.221), and *QoL* (Hedge's *g*: 0.09, Cl: -0.05 to 0.24, *p* (two-tailed): 0.215), from the study.

The authors concluded that no general effectiveness of EWI in cancer patients and survivors was found, psychological or physical. However, the study reasoned that, although modest, effects for subgroups of patients could be clinically relevant and suggested further study.

Study 2 – O'Toole et al, 2016, Cognitive behavioural therapies for informal caregivers of patients with cancer and cancer survivors - a systematic review and meta-analysis

The aim of the study was to examine the effect of Cognitive Behavioral Therapy (CBT) on psychological and physical health outcomes in informal caregivers (IC). The meta-analysis focused

on the following outcomes: mastery[7], psychological well-being, interpersonal well-being, physical well-being, and generic quality of life (global measures that could not be categorized as either psychological, physical or interpersonal). The meta-analysis included studies that (a) had a study population of informal caregivers of adult cancer patients or survivors, (b), included at least one quantitative measure of psychological, physical or interpersonal wellbeing at pre- and postintervention, (c) reported results that could be converted into an effect size, and (d) were written in English. The characteristics and data of the included studies are shown in table 8 (appendix).

The outcome results in this paper were Mastery (Hedge's *g*: 0.07, Cl: -0.02 to 0.16, *p*: 0.138), Psychological well-being (Hedge's *g*: 0.16, Cl: 0.07 to 0.24, *p*: < 0.001), Physical well-being (Hedge's *g*: 0.13, Cl: -0.01 to 0.17, *p*: 0.012), Interpersonal well-being (Hedge's *g*: 0.13, Cl: 0.04 to 0.22, *p*: 0.006), and QoL (Hedge's *g*: 0.02, Cl: -0.20 to 0.24, *p*: 0.868). All of these outcomes will be reanalyzed in this paper.

The authors concluded that only negligible effects of CBT on ICs were found and suggest that further studies focus on the basic affective sciences, to better understand and treat the emotional struggles of informal caregivers.

To highlight the package, the analysis of this paper solely uses the bhlm-package functions to analyse and produce interpretable results. Both examples are proof of concept, and for a full statistical analysis, the priors used in the examples would need more development based on prior knowledge of the field.

Data

First, the data was reviewed (see table 5 and table 8, appendix) to see if there were any in-regularities in the data frame that would cause issues to the bhlm-function. This meant that there needed to be a double grouped data setup (the two grouping factors, study*outcome number). There also needed to be an estimate column. For this example, Hedge's *g* effect size was chosen, but the package should work with any effect. The pre-processing done by the package, does not

[7] Mastery refers to appraisal efforts, self efficacy, coping skills, knowledge about cancer, and ability to perform caregiver related tasks of assisting the patient.

take missing values into account, as this should be done deliberately by the researcher and ensures that the researcher knows what data is sent to the model. Therefore, missing data rows were removed for the estimate column (none were found in the data for this study). Finally, a factorised outcome category column is needed for filtering chosen latent variables from the data.

Priors

For latent variable priors we chose "vague" gaussian priors (precision of 1), centered around zero. This prior will allow us to see whether the data will influence the posterior enough to "pull" the effect above zero. This choice reflects the research question of whether there is an effect different from 0 from the latent variables. JAGS estimates the expected predictive error (DIC) and effective number of parameters (pD) (Plummer, 2008). These approximations are widely used for Bayesian model comparison and could be used if we were to develop the model priors more (e.g. with bias correction or inclusion of observed variation, discussed later in the paper). Importantly, they do require that the effective number of parameters is much less than the number of samples. For theta a gaussian priors was chosen with mean 0 and precision 1. Due to time restriction, no bias was introduced to this prior, although future study would probably benefit from the introduction of effect size censoring bias (Guan & Vandekerckhove, 2016) as will be discussed later in the paper. For lambda, a gamma distribution prior was chosen which ensures that we get no negative values and are biased towards very small values. This is the most common distribution used for precision, ensuring that we have a conjugate prior that results in a wide sampling distribution.

Bayesian hypothesis testing

It's important to underline one conceptual problem to Bayesian hypothesis testing posed by Wagenmakers (2010). Wagenmakers argues that Bayesian hypothesis testing requires the researcher to be very wary of the "objectiveness" of the prior as "when the vagueness of the prior is arbitrary, so are the results from Bayesian hypothesis test". The priors chosen here could seem to risk falling victim to this arbitrariness, which often means that the Savage-Dickey ratio prefers the simpler model (e.g. a mean at zero). Wagenmakers poses several methods to increase the robustness of the Bayesian hypothesis testing. These include the local Bayes factor, the intrinsic Bayes factor, the fractional Bayes factor, and the partial Bayes factor. To further develop the package as a hypothesis testing tool, all these methods should be researched reviewed for inclusion in the method. These facts should be considered before concluding on the evidence of the posterior.

The weight needed for the Bayes factor is highly debated, but commonly a value above 3.3 is considered strong evidence and 5.0 is considered very strong (Jeffreys, 1961, p. 432)[8]. However, Alexander Etz and Joachim Vandekerckhove (2016), suggest a Bayes factor above 10 is suggested for decisive evidence. As it is most common, I will use the scale used by Jeffreys in the analysis.

Results

Study 1 – EWI

Table 1: EWI study BHLM model results.

Outcome	MAP	HPD[a]	CP[b]	BF01[c]	n.eff	Rhat	DIC[d]	pD[e]
Physical	0.055	[-0.33 : 0.42]	.409	4.441	1100	1.000		
Psychological	0.047	[-0.32 : 0.43]	.414	4.328	1100	1.001	43.1	22.0
QoL	0.021	[-0.39 : 0.48]	.455	3.949	1100	1.000		

[a]Highest posterior density or credible interval.
[b]Cumulated probability at null hypothesis (point of interest = 0).
[c]Bayes factor at null (Savage-Dickey ratio).
[d]DIC is an estimate of expected predictive error (lower is better) returned from JAGS.
[e]pD = $var(deviance/2)$, returned by JAGS.

The model seems to have converged with a Rhat below 1.00 for all outcomes and full effective sampling. The trace plots show no unhealthy traits in mixing or stationarity (Notice that with a thinning rate of 7 we have only kept every 7th sample).

Table 1 and 3: For the *Physical* (MAP = 0.055, HPD: -.33/.42, BF01 = 4.4) we see strong evidence for the posterior and against the prior. However, this new distribution only has a higher likelihood for the null effect and the distribution suggests a very small effect. Furthermore. with a cumulated probability of 41% below zero, we cannot disprove the null, but strengthen it instead. This result is very similar to the results shown by the original study (*g* = .08, p = .221). For the *Psychological* (MAP = 0.047, HPD: -.32/.43, BF01 = 4.3) and *QoL* (MAP = .021, HPD: -.39/.48, BF01 = 3.9). Although there is strong evidence against the prior the posterior distributions again suggest a very small effect and a higher likelihood for the null, and a cumulated probability of 41-45% below zero. Again, these results cannot disprove the null, and only strengthens the null hypothesis. These results are very

[8] *H. Jeffreys (1961). The Theory of Probability (3rd ed.). Oxford.* p. 432

similar to the results shown by the original study (Psychological combined: *g* = .04, p = .419) even though QoL shows an even smaller effect than the original (QoL: *g* = .09, p = .215).

Combined, the results of the BHLM analysis gives us very much the same results, as the ones seen in the original study which suggests very small inconclusive effects of EWI on Physical health, Psychological health, or Quality of Life. Furthermore, the Bayesian perspective allows us to see that the null hypothesis is strengthened in the posterior distributions, as the Bayes factor is above 1, indicating that an effect of zero is more likely under the posterior.

Study 2 – CBT

For the data, there was added a factor column for outcome names, as the data only had numbers.

```
dat = mutate(dat, Outcome2 = as.factor(Outcome)) %>%
  mutate(Outcome2 = fct_recode(Outcome, "Psychological" = "1", "Interpersonal" = "2",
                               "Physical" = "3","Mastery" = "4","QoL" = "5"))
```

Table 2: CBT study BHLM model results.

Outcome	MAP	HPD	CP	BF01	n.eff	Rhat	DIC	pD
Psychological	0.110	[-0.07 : 0.49]	.265	4.889	1100	1.002		
Interpersonal	0.177	[-0.17 : 0.36]	.159	3.428	1100	1.001		
Physical	0.273	[-0.37 : 0.34]	.114	2.501	1100	1.001	351.3	45.4
Mastery	0.155	[-0.12 : 0.46]	.181	4.024	1100	1.000		
Generic QoL	0.000	[-0.13 : 0.42]	.504	4.957	1100	1.001		

Savage-Dickey plots and trace plots for model one and two are shown in Appendix, table 6 and 7, respectively.

Table 2 and 5: The model seems to have converged with full effective samples and Rhat below 1.00. The trace plots show no unhealthy traits along the sampling process.

For *Psychological* (MAP = .11, HPD: -.07/.49, BF01 = 4.9), *Interpersonal* (MAP = .177, HPD: -.17/.36, BF01 = 3.4), and *Mastery* (MAP = .155, HPD: -.12/.46, BF01 = 3.4) we see strong evidence against the prior with all Bayes factors above 3.3. The distributions show much the same effects for Psychological health and Interpersonal as seen in the original study (Psychological: *g* = .16, p < .001, IP: *g* = 0.13, p = .006). However, the effects are still not conclusive with cumulated probabilities at 0 at 27% and 16%, respectively. To decisively disprove the null, a cumulated probability below 5% should be expected. More data to gain a more precise posterior, could maybe drive the results in this direction, but for now the results are inconclusive. For *Mastery* we actually see a stronger effect than shown in the original study (*g* = .07, p = .138), however, the results are still inconclusive with

CP at 18%. For *Physical* (MAP = .273, HPD: -.37/.34, BF01 = 2.5) we see that the Bayes factor suggests only substantial evidence against the prior. From the Savage-Dickey plot (Table 5) we see that this is caused by the fact that we simple do not have a big enough change in likelihood[9]. However, from the plot, we can see that, given more data, the posterior could become more precise and give us a SD ratio below 1. Further data collection would be needed on this variable to be sure. For *Generic QoL* (MAP = .000, HPD: -.13/.42, BF01 = 4.9) the effect is smallest and centered around the null.

Like with study 1, all effects strengthen the null hypothesis, as likelihood has increased on all outcomes at 0. Again, the results from the BHLM analysis can make the same conclusions as the original paper with negligible effects and the Bayesian perspective allowing us to see that the null hypothesis is strengthened.

Discussion

Results

Taken together, the results of the analyses using the BHLM package's method show very similar results to those reported in the original papers, providing initial support for the validity of the BHLM package. Given that both papers yielded relatively clear results, indicating negligible effects of the interventions analysed, this is to be expected. However, Bayesian inference allows for a more elaborate modelling in the form of the prior distribution than the one used in the present paper. Only "vague" (uninformative) priors were used in the present paper and it should be noted that the examples presented here primarily represent a proof of concept. While these results, however, still provide some insight, with a greater knowledge of the field, the priors could be further developed, which in turn should create more meaningful results, both in terms of parameter estimation and hypothesis testing.

[9] The multi-modal property of the posterior for the physical outcome, could be a result of random sampling from the log-density estimation. Testing with different random seeds removed the effect. Therefore, and for the sake of time, this property is not pursued further.

Future development of the package

It can be argued that a major contributing factor to the "reproducibility crisis" in psychology (and other sciences) and publication bias in general, is the reliance on statistical significance. Here, the Bayesian perspective provides results based on an analysis of the field and good choices in the prior distribution. It should be a positive development in research and publishing practice to weigh strength of analysis, argumentation, and interpretation of the prior distributions more than focusing on whether or not a significant result was achieved. In this way, this new method could be a tool to battle the issues currently afflicting most fields in science. This does not mean that current frequentist methods should be disregarded, as they already have developed several methods to counter the influence of publication bias, as seen in the analyses in the published example studies. However, the BHLM method could contribute to making the so far underutilized Bayesian methods more prominent in meta-analytical studies.

The package already pushes the user to focus on defining all priors for outcomes. In the future, the package should be developed to push the development of the prior distribution. This would mean introducing bias correction in some form. There are several methods that could be included in the model here. First, there is the bias correction method introduced by developed by Etz and Vanderkerckhove (2016). This type of bias correction could either be developed into the BHLM model, by having the priors subject to the four censoring functions or developed as a separate function to re-calculate the effect sizes, as demonstrated by Etz (2017). Ideally, the implementation of this method should report for all different censoring functions to give an image of all levels of assumed bias. Second, Eberly and Casella (1998) developed a Bayesian hierarchical model to estimate the number of unseen (unpublished) studies in a meta-analysis. Inclusion of this method could serve as both a good interpreting factor for the results shown by the Bayesian model.

Another aspect to be included in future versions of the model, would be the possibility of having different hierarchical level priors for each study included in the analysis. This should allow the user to have more control over the effect of studies on the latent variables. Users could use different levels of quality e.g., the Jadad scale or risk of bias assessment (Higgins & Green, 2011) as a sign of for example effect bias. Thus, low quality studies could have priors defined with positive skew or an increased mean to. This would decrease the latent variable's influence on the effect and in turn lower effect of low quality studies on the latent variables.

Currently, the model does not require standard deviations and standard error, but instead you define a prior through Lambda in the model. In the examples above, a common vague prior, biased towards low precision (wide distribution) was used. In the future builds, it might be viable to include the observed variance from studies. This could add additional meaningful information to the model from the data. A simple implementation with standard deviation is shown in figure 7, appendix.

The BHLM function currently has included the option to include another hierarchical level in the form of a "field" wide mean. However, one should use this option only when one has clear argumentation and reason for including such a prior. Further testing and research is needed to know whether the option is viable for inclusion in the method in future builds.

Conclusion

The BHLM package is an implementation of a new way to do meta-analysis with Bayesian inference and hierarchical latent mixture modelling. Still, all the ideas and prospects discussed above show that the package and method have much space for improvement. The results from the two example studies work as a proof of concept for the method and show that Bayesian methods can work side by side with the frequentist as they seem to give the same overall results, but the Bayesian methods allow for a more conceptual and wider analysis in the choosing of priors (both for outcomes, theta and lambda). Ideally, Bayesian methods would allow the scientific to move away from relying on statistical significance to relying on well-developed priors and interpretations of posterior probability distributions could proof as a solution to the issues afflicting psychological sciences today.

References

Bartlema, A., Lee, M., Wetzels, R., & Vanpaemel, W. (2014). *A Bayesian hierarchical mixture approach to individual differences: Case studies in selective attention and representation in category learning*. Journal of Mathematical Psychology, 59, 132-150.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013). *Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation*. Journal of biopharmaceutical statistics, 23 (6), 1352-1371.

Charles Kooperberg (2015). *polspline: Polynomial Spline Routines*. R package version 1.1.12. https://CRAN.R-project.org/package=polspline

Duval S. & Tweedie R. *Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in metaanalysis.* Biometrics 2000;56 (2): 455–463.

Dickey, J. M., & Lientz, B. P. (1970). *The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain.* The Annals of Mathematical Statistics, 41, 214–226.

Deng, H., & Wickham, H. (2011). *Density estimation in R.* Electronic publication.

Eberly, L. E., & Casella, G. (1999). *Bayesian estimation of the number of unseen studies in a meta-analysis.* Journal of Official Statistics, 15 (4), 477.

Edwards, W., Lindman, H., & Savage, L. J. (1963). *Bayesian statistical inference for psychological research*. Psychological Review, 70, 193–242.

Etz, A. (2017, May 21) *Slides: Bayesian Bias Correction: Critically evaluating sets of studies in the presence of publication bias*. Retrieved from https://alexanderetz.com/2017/05/21/slides-bayesian-bias-correction-critically-evaluating-sets-of-studies-in-the-presence-of-publication-bias/

Etz A. & Vandekerckhove J. (2016). *A Bayesian Perspective on the Reproducibility Project: Psychology*. PLoS ONE 11(2): e0149794.

Gelman, A. & Rubin, D. (1992). *Inference from iterative simulation using multiple sequences*. Statistical Science, 7:457–511.

Glass, G. V. (1976). *Primary, secondary, and meta-analysis of research*. Educational researcher, 5 (10), 3-8.

Glass, G. V. (2015). *Meta-analysis at middle age: a personal history*. Research synthesis methods, 6 (3), 221-231.

Grant R. L. (2018). *The uptake of Bayesian methods in biomedical meta-analyses: a scoping review, 2005-2016*. Pre-print.

Guan, M., & Vandekerckhove, J. (2016). *A Bayesian approach to mitigation of publication bias.* Psychonomic bulletin & review, 23 (1), 74-86.

Haidich, A. B. (2010). *Meta-analysis in medical research*. Hippokratia, 14 (Suppl 1), 29-37.

Ioannidis, J. P. (2005). *Why most published research findings are false*. PLoS medicine, 2 (8), e124.

Higgins, J. P., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. 4). John Wiley & Sons.

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). *Evidence based medicine: The case of the misleading funnel plot.* BMJ: British Medical Journal, 333 (7568), 597.

Light, R. J. & Pillemer, D. B. (1984). *Summing up. The science of reviewing research*. Cambridge, MA: Harvard University Press.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan (Vol. 122)*. CRC Press.

Masicampo, E. J., & Lalande, D. R. (2012). *A peculiar prevalence of p values just below .05*. Quarterly journal of experimental psychology, 65(11), 2271-2279.

McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy*. Yale University Press.

Open Science Collaboration. (2015). *Estimating the reproducibility of psychological science*. Science, 349 (6251), aac4716.

O'toole, M. S., Zachariae, R., Renna, M. E., Mennin, D. S., & Applebaum, A. (2017). *Cognitive behavioral therapies for informal caregivers of patients with cancer and cancer survivors: a systematic review and meta-analysis*. Psycho-oncology, 26 (4), 428-437.

Pennebaker JW, Beall SK. *Confronting a traumatic event: toward an understanding of inhibition and disease*. J abnorm Psychol 1986;95 (3):274–281.

Plummer, M. (2008). *Penalized loss functions for Bayesian model comparison*. Biostatistics, 9 (3), 523-539.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Scott, D. W., R port by Gebhardt, A., and adopted to recent S-PLUS by Kaluzny, S. (2015). *ash: David Scott's ASH Routines*. R package version 1.0-15. https://CRAN.R-project.org/package=ash

Stan Development Team. (2018). *RStan: the R interface to Stan.* R package version 2.17. 3. http://mc-stan.org/

Su, Y. S., & Yajima, M. (2012). *R2jags: A Package for Running jags from R*. R package version 0.05-7, http://CRAN.R-project.org/package=R2jags

Wagenmakers, E. J. (2007). *A practical solution to the pervasive problems of p values.* Psychonomic bulletin & review, 14 (5), 779-804.

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

Wand, M. (2015). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*. R package version 2.23-15. https://CRAN.R-project.org/package=KernSmooth

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

Zachariae, R., & O'toole, M. S. (2015). *The effect of expressive writing intervention on psychological and physical health outcomes in cancer patients—a systematic review and meta-analysis*. Psycho-Oncology, 24 (11), 1349-1359.

Appendix

Study 1 – EWI

Table 3: EWI study model 1 (precision: 1). Savage-Dickey and Trace plots (Iteration thinning rate: 7)

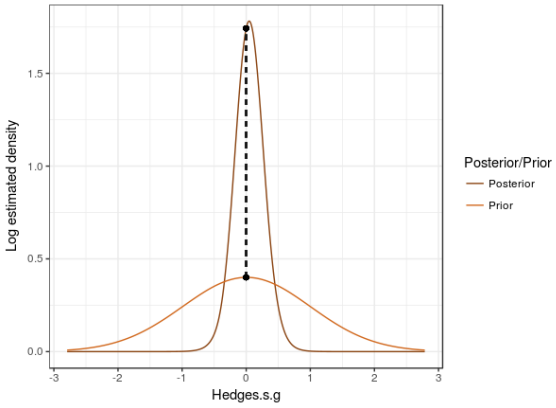
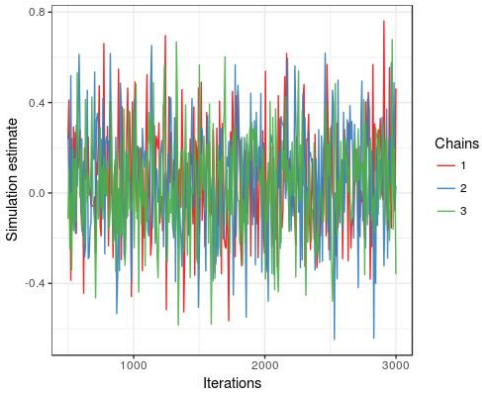
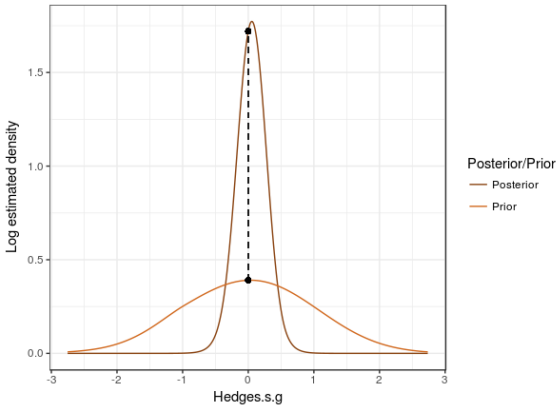
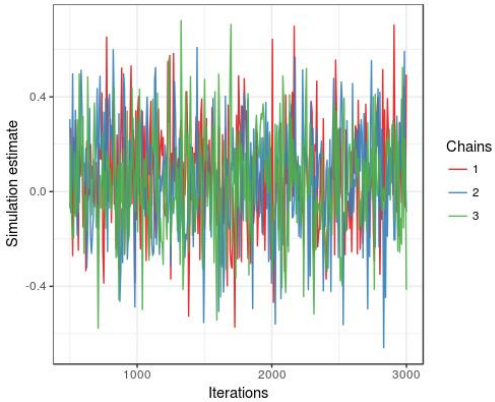
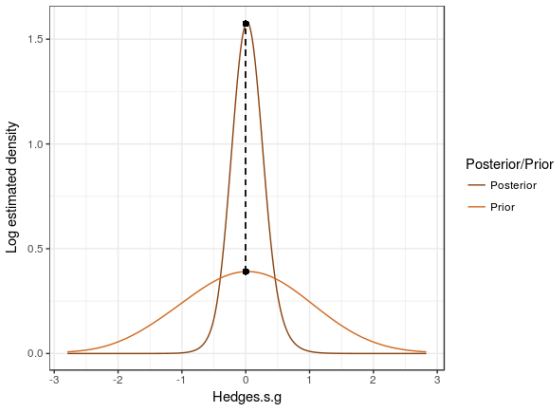
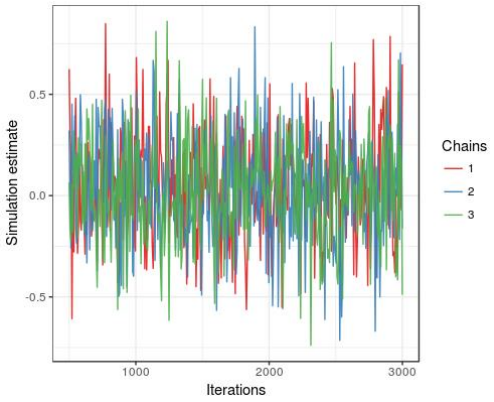
Outcome	Savage-Dickey plot	Trace plot
Psychological		
Physical		
Quality of Life		

Table 4: EWI meta-analysis, data (head)a.

Study	Study number[b]	Outcome number	Hedges's *g*	SE	Measures	Outcome[c]
Walker et al. 1999	1	1	-0.269	0.346	Avoidance	Psychological
Walker et al. 1999	1	2	-0.154	0.345	Intrusive thoughts	Psychological
Walker et al. 1999	1	3	-0.212	0.346	Negative mood	Psychological
Walker et al. 1999	1	4	-0.013	0.345	Positive mood	Psychological
de Moor et al. 2002	2	1	-0.362	0.351	Avoidance	Psychological
de Moor et al. 2002	2	2	-0.068	0.346	Intrusive thoughts	Psychological
de Moor et al. 2002	2	3	0.214	0.348	Negative mood	Psychological
de Moor et al. 2002	2	4	0.167	0.347	Perceived stress	Psychological
de Moor et al. 2002	2	5	0.714	0.368	Sleep	Physical
Rosenberg et al. 2002	3	1	0.446	0.360	Health contacts	Physical
(…)	(…)	(…)	(…)	(…)	(…)	(…)

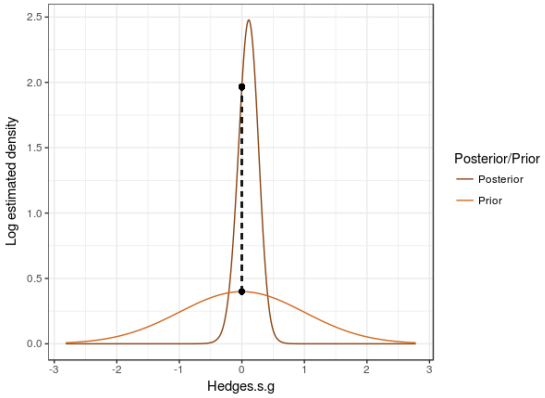
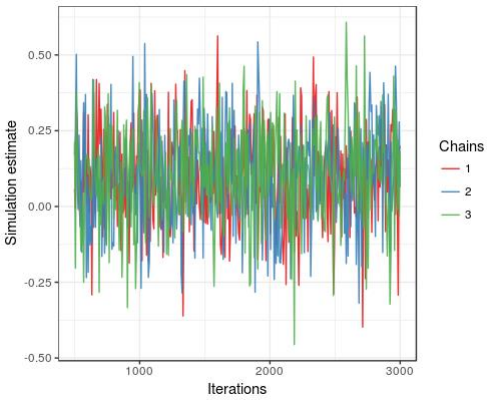
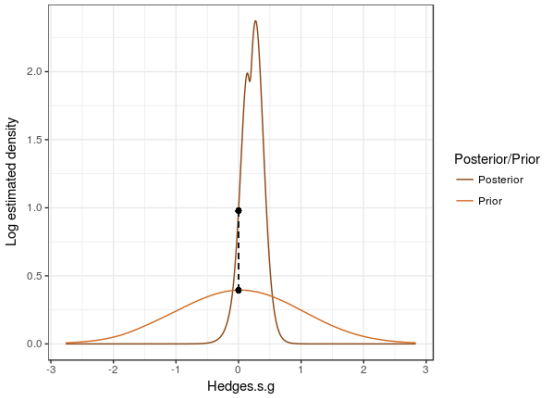
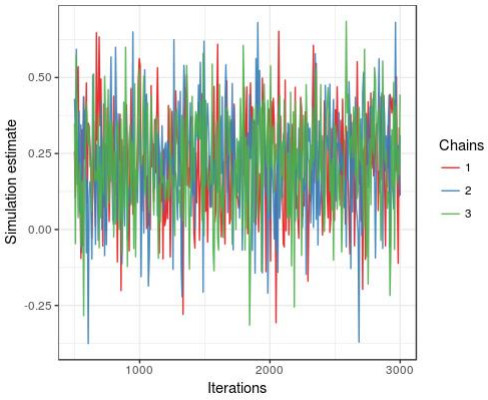
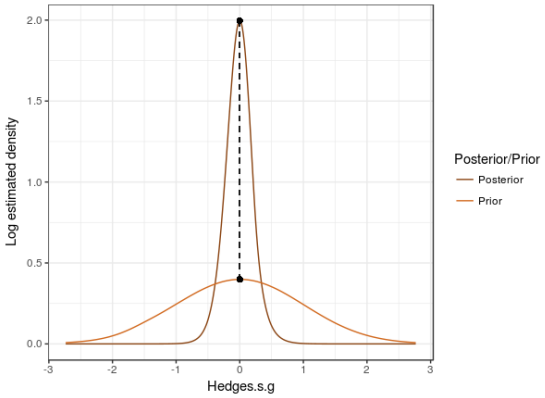
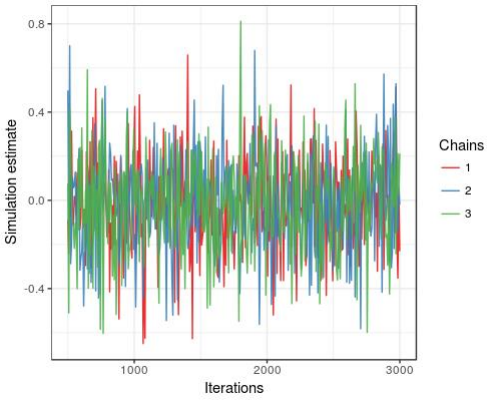
[a] Variables irrelevant to the analysis of the present paper are omitted
[b]15 studies with a mean of 4.2 outcomes per study (high: 7, low: 2).
[c]Outcome categories. 21 Physical (from 10 studies), 34 Psychological (from 13 studies), and 8 QoL (from 5 studies).

```
model{

  lambda~dgamma(0.001, 0.001)

  Physical~dnorm(0, 1)
  Psychological~dnorm(0, 1)
  QoL~dnorm(0, 1)

  outcome_options[1] <- Physical
  outcome_options[2] <- Psychological
  outcome_options[3] <- QoL

  for (s in 1:upper_group) {

    theta[s]~dnorm(0,1)

    for (o in start_bounds[s]:(start_bounds[s+1]-1)) {

      outcome[s,o] <-  outcome_options[outcomes_numeric[o]]
      eta[s,o] <- theta[s] + outcome[s,o]
      outcomes[o] ~ dnorm(eta[s,o],lambda)

    }
  }
}
```

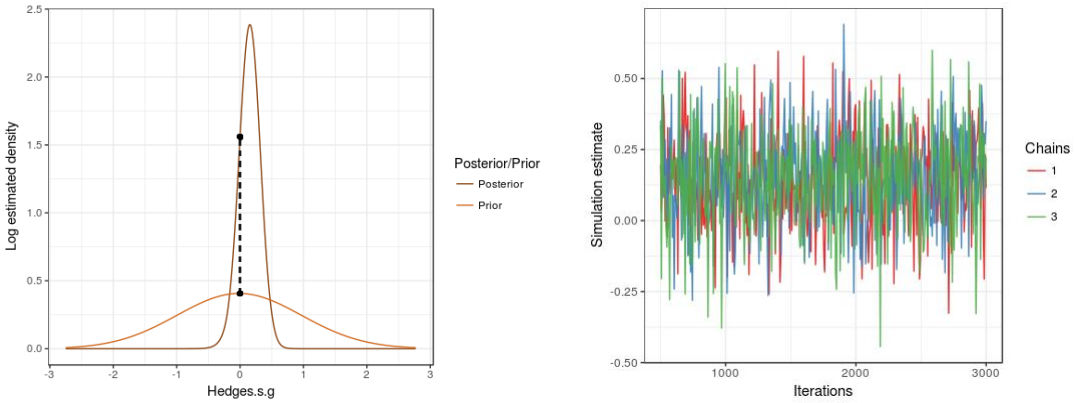
Figure 5: EWI study, model 1 implementation

Study 2 – CBT

Table 5: CBT study model 1 (precision: 1). Savage-Dickey and Trace plots (Iteration thinning rate: 7)

Outcome	Savage-Dickey plot	Trace plot
Psychological		
Physical		
Quality of Life		

Mastery



Interpersonal

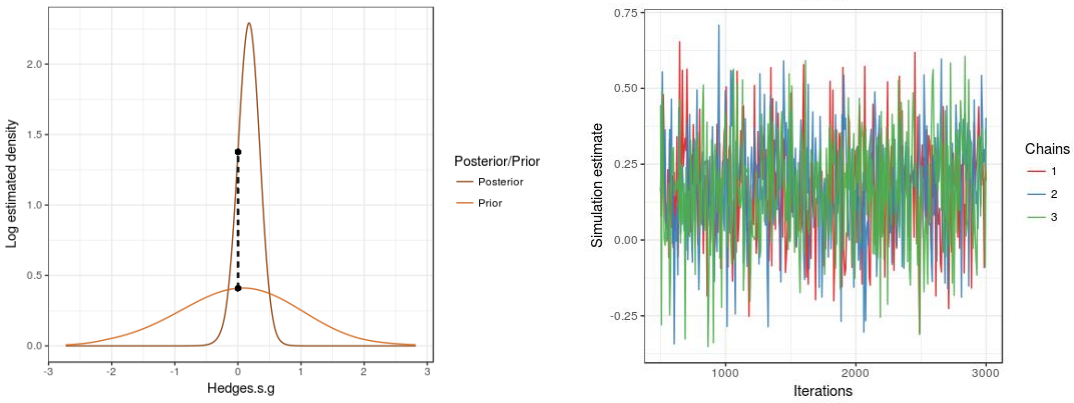


Table 6: CBT meta-analysis, data (head).

Study	**Study number**[a]	**Outcome number**	**Hedge's *g***	**SE**	**RCT/OT**	**Outcome**[b]
Baucom et al., 2009	1	1	-0.655	0.521	RCT	1 (Psychological)
Baucom et al., 2009	1	2	0.281	0.508	RCT	1 (Psychological)
Baucom et al., 2009	1	3	0.506	0.515	RCT	2 (Interpersonal)
Baucom et al., 2009	1	4	0.369	0.510	RCT	3 (Physical)
Bevans et al., 2010	2	1	0.068	0.290	OT	4 (Mastery)
Bevans et al., 2010	2	2	-0.202	0.293	OT	1 (Psychological)
Bevans et al., 2010	2	3	0.348	0.299	OT	3 (Physical)
Bevans et al., 2013	3	1	0.389	0.127	OT	1 (Psychological)
Bevans et al., 2013	3	2	0.266	0.125	OT	1 (Psychological)
Birnie et al., 2010	4	1	0.488	0.252	OT	4 (Mastery)
(…)	(…)	(…)	(…)	(…)	(…)	(…)

[a]36 studies with a mean of 6.528 outcomes per study (high: 28, low:1).
[b]The five outcome categories. 93 Psychological well-being (from 31 studies), 36 Interpersonal well-being (from 16 studies), 37 Physical well-being (from 18 studies), 57 Mastery (from 16 studies), and 12 Quality of Life (from 10 studies). The data received, had only numbers, but cross-check with studies of the meta-analysis, found the correct outcome names.

```
model{

  lambda~dgamma(0.001, 0.001)

  Psychological~dnorm(0, 1)
  Interpersonal~dnorm(0, 1)
  Physical~dnorm(0, 1)
  Mastery~dnorm(0, 1)
  QoL~dnorm(0, 1)

  outcome_options[1] <- Psychological
  outcome_options[2] <- Interpersonal
  outcome_options[3] <- Physical
  outcome_options[4] <- Mastery
  outcome_options[5] <- QoL

  for (s in 1:upper_group) {

    theta[s]~dnorm(0,1)

    for (o in start_bounds[s]:(start_bounds[s+1]-1)) {

      outcome[s,o] <-  outcome_options[outcomes_numeric[o]]
      eta[s,o] <- theta[s] + outcome[s,o]
      outcomes[o] ~ dnorm(eta[s,o],lambda)

    }
  }
}
```

Figure 6: CBT study, model implementation.

Additional

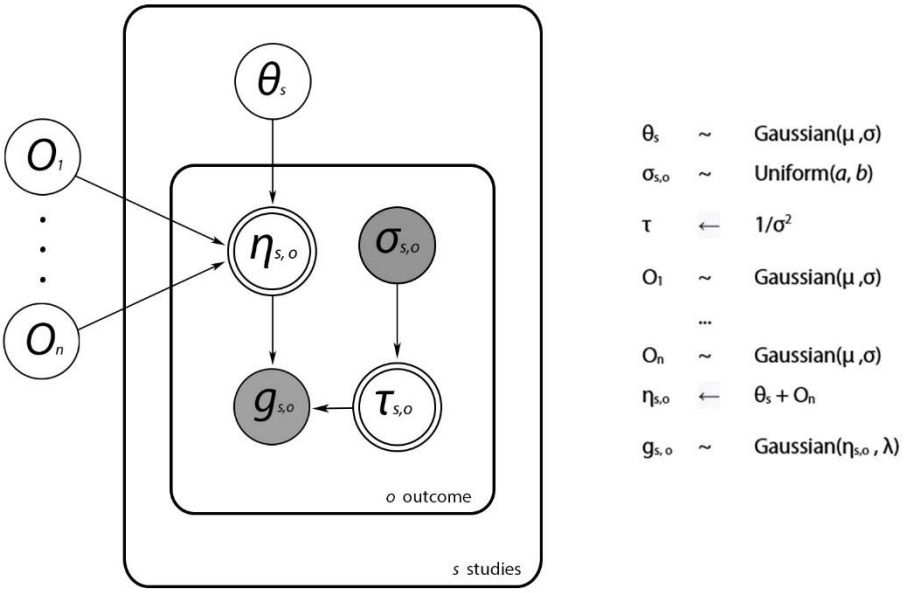


Figure 7: Model with observed standard deviation.