Bayesian Analysis of Randomized Controlled Trials

Julian Bautista, Alex Pavlakis, Advait Rajagopal

The New School for Social Research, 6 E 16 St, New York, NY, 10003

July 23, 2018

Short Running Title: Bayesian Randomized Controlled Trials

Abstract

Objective : This paper is an introduction to Bayesian data analysis for empirical researchers in the field of clinical psychology, focusing on applications for the study of eating disorders. We summarize the intuition and methodology of Bayesian data analysis, and motivate its use for analyzing Randomized Controlled Trials (RCT). We demonstrate the strengths of the approach through the analysis of a Cognitive Behavioral Therapy RCT on the influence of a smartphone application on binge-eating disorder (BED) and bulimia nervosa.

Method : We fit a multilevel Poisson regression model on outcome variable Objective Bulimic Episodes (OBE) as a function of the treatment and other covariates. OBE is a discrete count variable and the Poisson model fits well. The multilevel structure accounts for individual and time-varying effects.

Results : Our analysis suggests that the smartphone application causes a reduction in the instances of OBE for patients in the initial weeks but the effect may wear off by the end of the treatment period. Bayesian methods allow us to obtain heterogenous treatment effects for different individuals and stages of the therapy while explicitly modeling uncertainty around these effects.

Discussion : We conclude that Bayesian methods are a powerful tool for incorporating data and prior information into models. They have the potential to improve analyses of RCTs in eating disorders given small sample sizes, small effect size, abundant prior information, heterogenous participants, and experimental design. These methods are useful for empirical researchers, particularly clinical psychologists.

Keywords – Methods, Randomized Controlled Trial, Bayesian Analysis, Eating Disorders, binge-eating Disorders, Statistics

Abstract word count – 237

Main text word count – 3,958

Acknowledgements – The authors would like to thank Tom Hildebrandt, Andreas Michaelides, Diana Mackinnon, Rebecca Greif, Lynn DeBar, and Robyn Sysko, who designed, managed, and analyzed the randomized controlled trial referenced in this paper. Tom Hildebrandt also provided crucial guidance as the paper was being written; it would not have been possible without him. We also want to thank Prof. Andrew Gelman for connecting us with this project and his contribution to Bayesian statistics.

Funding – R41-MH096435; PI: Hildebrandt

1 Introduction

Bayesian data analysis (BDA) is a method for building statistical models to describe data. Researchers begin with explicit assumptions about the data generating process based on past research and the scientific nature of the problem. Based on these beliefs and assumptions, they then collect data from designed experiments or observational studies. Using Bayes' rule they combine their assumptions or *prior information* with the actual data into a comprehensive probability model. This model contains information about all known (observed data) and unknown (unobserved parameters) quantities related to the data generating process. Bayesian data analysis is gaining popularity in medical, pharmaceutical, and social-science research because it allows researchers to combine prior information with data to model data generating processes.

The purpose of this paper is to motivate the use of BDA in applied research, especially for analyzing results of Randomized Controlled Trials (RCT). The standard models for analyzing RCTs are regression models because they allow researchers to estimate causal effects precisely by controlling for available pre-treatment covariates (Gelman & Hill, 2006). The Bayesian approach to regression modeling allows researchers to build statistical models which include relevant prior information, to estimate varying treatment effects for subgroups of the population, and to model the uncertainty in their estimates explicitly. These improvements in the analysis of treatment effects make Bayesian methods especially valuable for empirical researchers in psychiatry and clinical psychology. In this paper, we use data from a recently published RCT by Hildebrandt et al. (2017) examining the efficacy of a smartphone application designed to augment the efficacy of cognitive behavioral therapy guided self-help (CBT-GSH) treatment for reducing binge-eating episodes.

This paper is an introduction to the Bayesian approach to analyze RCT data. We focus on providing motivation for the usefulness of the approach and guidelines for executing it in a rigorous way, particularly in the context of treating eating disorders. We model data used by Hildebrandt et al. (2017) using a multilevel/hierarchical Poisson model (explained in Results section). We use a discrete distribution (Poisson) to account for the non-negative and discrete outcome variable (OBE), that contains a large number of 0's. The innovation in using the Bayesian approach is that we are able to control for pre-treatment covariates while capturing heterogeneity in treatment effects across individuals and over time and explicitly modeling the uncertainty around those effects. We find that we are able to capture the pattern of excess 0's and positive skew in the data and our model predicts the data well.

2 Methods

2.1 Bayesian Data Analysis

There are three main steps to BDA which are listed below and explained in detail in the subsequent sections:

1. Specify a model based on scientific knowledge of the data generating process.
2. Estimate model parameters based on observed data.
3. Evaluate the model's accuracy and expand or alter the model.

2.2 Probability notation and Bayes' rule

In Bayesian Data Analysis, assumptions are specified mathematically as *prior distributions.* Data are represented through a *likelihood model.* Bayes' Rule combines prior distribution and data likelihood into a *posterior distribution.* A formal expression of the Bayes' Rule is as follows;

$$p(\theta|y,x) = \frac{p(\theta)p(y|\theta,x)}{p(y)} \tag{1}$$

θ is the estimated parameter. In the context of an RCT, θ is the effect of the treatment x on dependent variable y. $p(\theta)$ is the prior distribution of the treatment effect, which captures the researcher's beliefs about the model parameter prior to any analysis. $p(y|\theta,x)$ is the likelihood function and is the probability of the observed data given the parameter. $p(y)$ is a normalizing constant with respect to θ that ensures the left hand side $p(\theta|y,x)$ is a proper probability distribution that integrates to 1. For a full treatment of proper distributions and normalizing constants see Gelman et al. (2014).

For our purpose, we can ignore this denominator and rewrite the expression for the posterior $p(\theta|y)$ as;

$$p(\theta|y,x) \propto p(\theta)p(y|\theta,x) \tag{2}$$

The posterior distribution is proportional to the product of the prior and the likelihood. Ultimately the goal of modeling is to learn the posterior distribution $p(\theta|y,x)$ and summarize it accurately (Gelman et al., 2014). Based on this we can make inferences and predictions.

Model development

Model development involves specifying a model that accounts for all observed data and unobserved parameters. The model should include all knowledge of the experiment or data collection process and should be logically consistent with scientific nature of the problem. We approach model development in three steps.

2.3 Exploratory data analysis

Summarizing data through visualizations and summary statistics is a critical model building step. We look at summary statistics of our relevant variables to understand their distributions. We explore data through univariate analyses, such as histograms to visualize distributions, and bivariate analyses, such as scatter plots to see if there are meaningful linear or non linear relationships. This exploratory process informs us of the distribution of individual variables, their correlations, and other relevant information for modeling.

Exploratory analysis helps us to choose an appropriate likelihood model to describe the data. For example, if our variables are normally distributed and linearly correlated, we might choose a linear regression model. If the relationship between variables is not straightforward we might choose more complicated models. It allows researchers to explore and solidify intuition about the problem at hand and the nature of the data. An introductory treatment of exploratory data analysis can be found in Tukey (1977). A more advanced explanation of exploratory data analysis in the context of Bayesian statistics is available in Gabry, Simpson, Vehtari, Betancourt, and Gelman (2017) and Betancourt (2018).

2.3.1 Setting up a likelihood model

The likelihood function is often analogous to a traditional regression equation. The researcher must select independent variables that represent important determinants of the outcome variable. The likelihood represents the distribution of the outcome variable given the independent variables and model parameters. To choose the right likelihood function, it is critical to know the type of data. Data can be binary, categorical, ordinal, count, or continuous and each of these types of data require a different kind of model. In the example in this paper, the outcome variable is non-negative and discrete so we choose a Poisson likelihood function (see Results section).

2.3.2 Choosing a prior distribution

The prior distribution is a mathematical encoding of researchers' assumptions. A prior distribution serves three functions. First, it makes assumptions about the underlying scientific nature of the problem explicit. Second, it regularizes or constrains the parameter space by specifying likely ranges for parameter values. Third, it facilitates the calculation of a posterior distribution and makes it possible to generate simulations from that distribution.

Prior choice depends on the parameter or coefficient of interest. We could assign a completely "noninformative" or flat prior to our coefficient by specifying a uniform distribution as the prior. This is equivalent to saying that our parameter is equally likely to assume any value from negative infinity to positive infinity and we have no more information about it. Using a noninformative uniform prior is the same as carrying out a maximum likelihood estimation of the parameter of interest.

Researchers rarely know nothing about the problem or relevant parameters. They often possess valuable information about the parameter which statisticians can incorporate as informative prior information. For instance, if a coefficient of interest is a proportion, then it *must* be between 0 and 1, and we can assign a *Beta* prior to that coefficient. If we believe that a coefficient is close to zero, but may be positive or negative, we could assign an informative $Normal(0, 1)$ prior distribution to it. More information about prior choice for the same can be found on the GitHub page for Stan developers (Stan Development Team, 2015). Some plots displaying common priors can be found in the Appendix. We explain the Bayesian probabilistic programming language, Stan, in model estimation.

2.4 Model estimation

Once the likelihood model and prior distribution are specified, the posterior distribution of an outcome variable can be estimated. As discussed earlier, the posterior distribution is obtained by multiplying the prior and likelihood. The distribution obtained by this process is proportional to the true posterior because we can ignore the normalizing constant (the denominator in equation 1). We use an approximation of the posterior as shown in equation 2 because calculating the true posterior analytically may be practically impossible. A standard practice is to use Markov Chain Monte Carlo (MCMC) sampling methods to approximate the posterior up to a normalizing constant and sample from it (Gelman et al., 2014). There are other approaches to calculating the posterior distribution, but those are beyond the scope of this paper. Analyses in this paper are carried out with the Bayesian probabilistic modeling language Stan. The R interface of the language can be understood at Stan Development Team (2018) and Carpenter et al. (2017) present a clear conceptual overview of the language. Stan uses a Hamiltonian Monte Carlo sampling algorithm (from a broader class of MCMC sampling methods) to approximate the posterior distribution. Betancourt (2017) has a clear exposition of how the algorithm works. Stan returns the full posterior distribution of the desired parameters. Stan can also predict values based on the specified model which can be used for model checking, validation, and expansion.

2.5 Model checking and expansion

We evaluate whether our model explains the data by investigating parameter distributions and posterior predictive checking. Visualizations of parameter distributions, such as histograms, enable us to summarize estimates and the uncertainty around them. Posterior predictive checks involve "simulating replicated data under the fitted model and then comparing these to the observed data" (Gelman & Hill, 2006, p. 158). While we may not expect our model to generate our data exactly, it should recover important patterns. Once we have estimated a model and studied its properties, we can *expand* it through reparameterizations, adding parameters, or changing prior distributions.

After we have fit multiple models, we may want to compare their performances. There is wide literature on the most effective ways to compare and assess Bayesian models. These include cross validation methods, Bayes factors, and information criterion. One popular Bayesian approach to model checking, is *leave-one-out-cross validation (loo-cv).* The process compares models by fitting them on all data points except one, then evaluating how they predict the remaining data point. This process is repeated until all data points have been left out once. There are also several information criterion that can be used. An exhaustive summary of these "practical" model checking methods can be found in Vehtari, Gelman, and Gabry (2017). Some model checking methods that are closer to the null-hypothesis significance testing (NHST) framework are the ROPE (region of practical equivalence)(Kruschke, 2014) and Bayes factor (Rouder, Speckman, Sun, Morey, & Iverson, 2009) approaches.

Model checking is a critical step in Bayesian Data Analysis; however, there is no one-size-fits-all approach. We recommend posterior predictive checks because it is a direct way to assess the model fit to various aspects of data. By using posterior predictive checks we neither "accept" nor "reject" models but aim to understand their limitations in realistic replications (Gelman et al., 2014).

2.6 Experiment: Impact of Smartphone App on Eating Disorder Behavior

Hildebrandt et al. (2017) conducted an experiment to test whether the Noom Monitor, a smartphone application, could augment the effect of in-person therapeutic treatment on binge-eating behavior. The treatment, known as *guided self-help treatments based on cognitive-behavior therapy* (CBT-GSH), had been shown in previous research to reduce binge-eating behavior by 10-50%. The Noom Monitor application was designed to facilitate CBT-GSH. For this example, we consider two research questions from the experiment:

1. Is CBT-GSH more effective at reducing binge-eating behavior when facilitated by the Noom Monitor?

2. Does the effect of the Noom Monitor vary over time?

2.7 Experimental design

66 men and women with Bulimia Nervosa (BN) or binge-eating Disorder (BED) were randomly assigned into two treatment conditions: CBT-GSH (N= 33) or CBT-GSH + Noom (N=33). Therapy lasted for 12 weeks. Assessments were conducted at weeks 0, 4, 8, 12, 24, and 36. The primary outcome was Objective Bulimic

Episodes (OBE). For more information about the experiment, choice of dependent (outcome) variable, historical context and past research in this area, see Hildebrandt et al. (2017). There is a discussion of Bulimia Nervosa and binge-eating as well as an explanation of the choice of outcome variable OBE. Hereafter we refer to CBT-GSH as the control condition/group and CBT-GSH + Noom as the treatment condition/group.

3 Results

3.1 Model development

3.2 Exploratory data analysis

We start the analysis by plotting the outcome variable OBE. Figure 1 displays OBEs per week for each individual in both treatment conditions. A few aspects of the data immediately stand out, which suggest that any model should account for individual-level effects and time-level effects, and should let treatment effects vary over time.

1. The number of OBEs decreases over the course of the treatment for almost all subjects.
2. The biggest decreases in OBEs appear to occur in the early stages of treatment.
3. The primary sources of variation in OBE appear to be *between people* and *over time*.

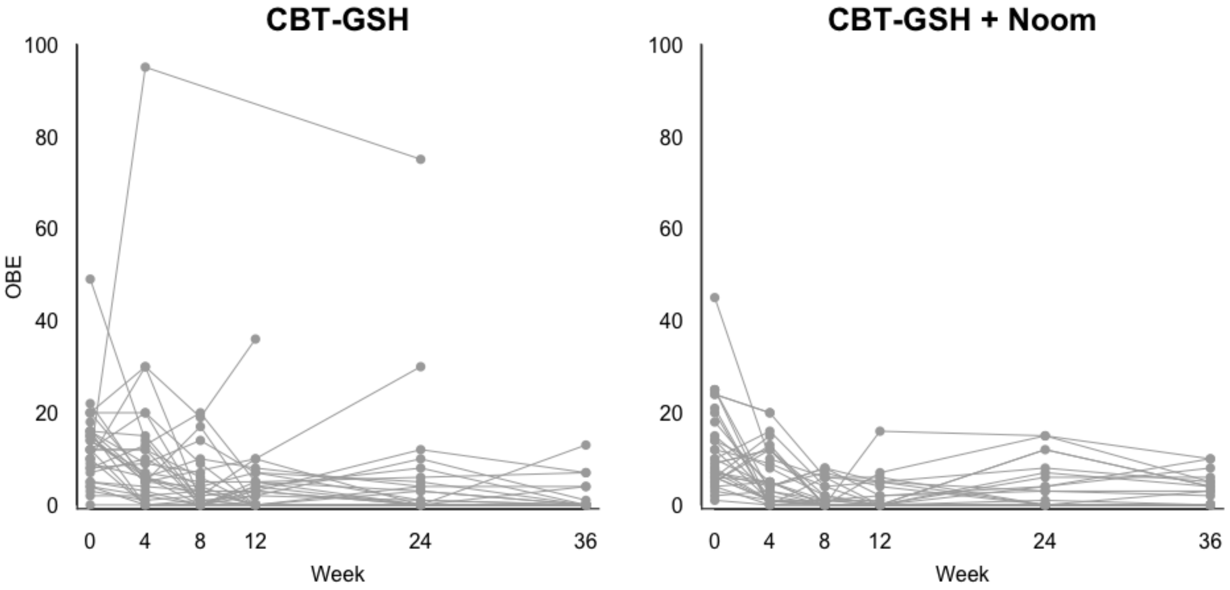

Figure 1: *This figure shows the OBE measurements for individuals divided into the treatment and control group. Fig.1 (a) and (b) show the OBE measurements for individuals in the CBT-GSH group and CBT-GSH + Noom group respectively. The horizontal axis shows the week and the vertical axis shows the instances of OBE. The gray dots represent OBE readings over time for each individual.*

Figure 2 displays the distribution of OBEs in each condition in each week, aggregating across individuals. We notice three characteristics of the data from these histograms.

1. The distributions appear to condense around zero for both conditions over time
2. The distributions in the CBT-GSH condition appear to have longer tails than those in the CBT-GSH+Noom condition
3. OBEs are count data; they must be nonnegative integers.

These three characteristics suggest that the appropriate model for OBEs is the Poisson distribution, because it is restricted to nonnegative integers and can concentrate its density around low numbers with a long tail.

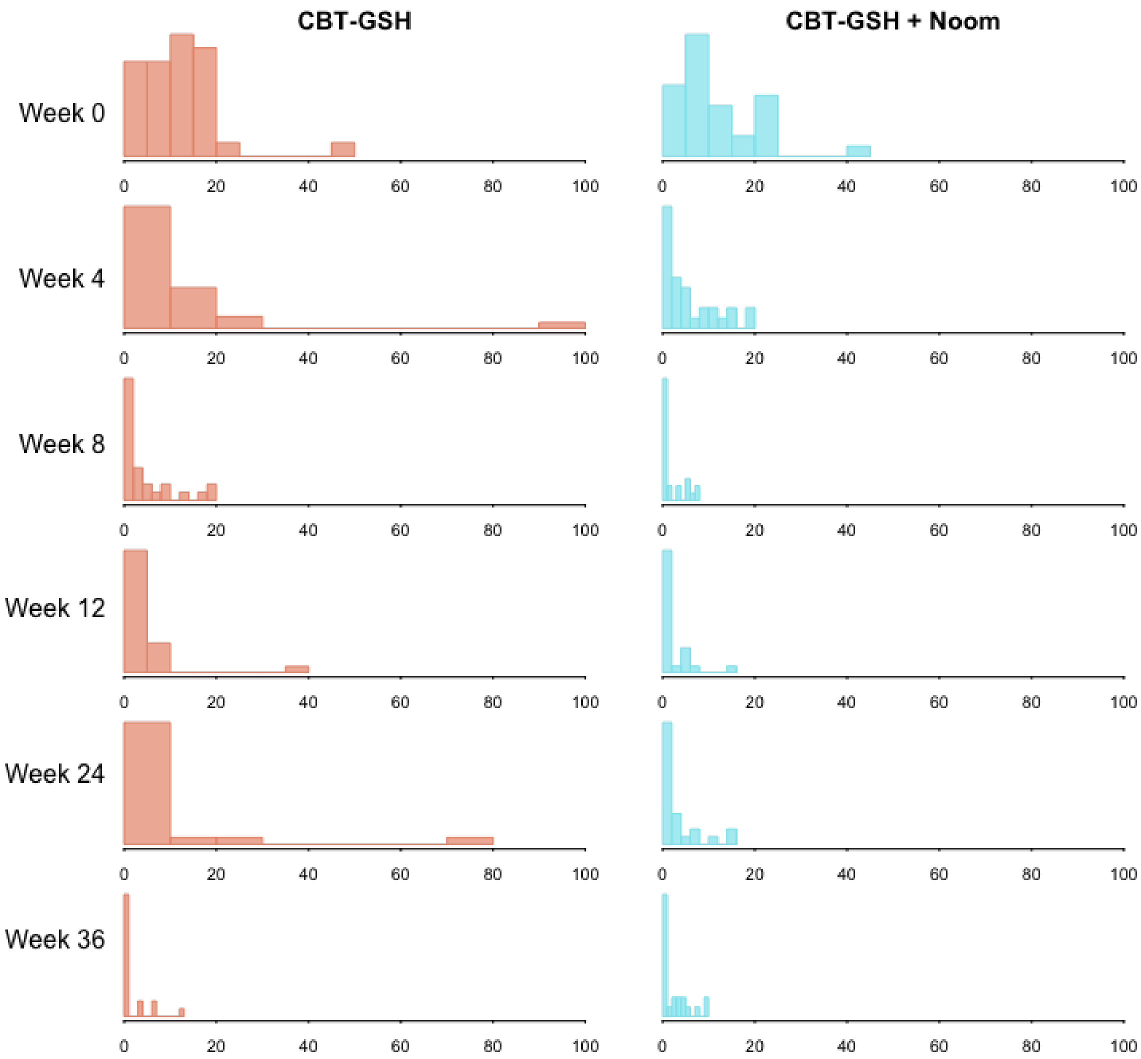

Figure 2: *The histograms display the distribution of OBEs in the control and treatment condition in each stage of the treatment. The orange and light blue histograms show the distribution of OBEs for the CBT-GSH and CBT-GSH + Noom group respectively.*

3.3 Setting up a likelihood

We analyze RCTs by modeling the outcome of interest (in this case OBE) as a function of the treatment and all available pre-treatment covariates. The coefficients associated with the treatment are estimates of average treatment effects. Inclusion of all available pre-treatment covariates accounts for variation in the outcome variable, decreasing uncertainly around treatment effects and providing the model with more predictive power. We conduct *intent-to-treat* analysis, meaning that our inferences will be based on initial treatment assignment, and will not account for mid-experiment dropouts.

The outcome variable is restricted to be nonnegative integers, so we fit a Poisson regression model, partially pooling across individuals, time periods, and treatment effects. For each individual in each time period, the number of OBEs follows a Poisson distribution, with a mean dependent on the characteristics of the individual and the time period.

An important implication of our likelihood function is that unlike previous zero-inflated Poisson (ZIP) and zero-inflated negative binomial models (Grotzinger, Hildebrandt, & Yu, 2015; Hildebrandt et al., 2017) for binge-eating data, we have a single generative process. We do not have a different generative process for 0's from abstinence, full remission and the non-0 count data. By accounting for individual level variation across time, in other words partially pooling (Gelman et al., 2014) across individuals and time periods, we are able to capture each individual's behavior and model their associated OBEs across time.

$$OBE_{i,t} \sim Poisson(\lambda_{i,t}) \tag{3}$$

$$\lambda_{i,t} = exp(\alpha_i + \beta_t + \gamma_t T_i + X_i \theta) \tag{4}$$

$$T_i = \begin{cases} 0, & \text{if } CBT - GSH \\ 1, & \text{if } CBT - GSH + Noom \end{cases} \tag{5}$$

α is an individual-specific intercept, β is a time-specific intercept, γ is a time-specific treatment effect, T is a treatment indicator, X is a matrix of individual level covariates (age, sex, race, etc), and θ is a vector of effects. Subscripts $i = 1, ..., 66$ indicate individuals and subscripts $t = 0, 4, 8, 12, 24, 36$ indicate time periods. We validate the effectiveness of our probability model using posterior predictive checks (Fig.5). We are able to capture the trend of inflated zeros for all time periods for both groups of individuals (treatment and control). Our model learns and reproduces the positive skewness and long tails of the OBE distribution.

3.4 Choosing a prior distribution

Table 1 has a list of different sources from which prior information has been obtained for this experiment. It aims to summarize the various methods which a researcher can use to incorporate prior information into the modeling process. There is a rich literature on binge-eating disorders and bulimia nervosa studies, implying a large amount of prior information. This makes analysis of similar RCTs amenable to Bayesian methods. For more examples of priors see the Appendix or check the "Prior Choice Recommendations" GitHub page (Stan Development Team, 2015).

Source of Prior Information	
Experimental Design	Outcome variable is nonnegative integers
Literature	Treatment effect size is small
Exploratory Data Analysis	A large number of zeros in OBE data due to remission
	There is variation in OBEs at the individual level
	There is variation in OBEs over time
	Treatment effects may vary over time

Table 1: *Sources of prior information.*

We believe that individual-level intercepts are simultaneously unique to the individual and common to the population; that is, each individual has their own baseline predilection to engage in eating disorder behavior, but those baseline predilections are not drastically different from each other. We operationalize this concept by modeling all individual-level intercepts as coming from a common distribution, with *hyperparameters* μ_α and τ_α. In Bayesian statistics, hyperparameters are parameters of prior distributions. In hierarchical models, we model hyperparameters explicitly.

$$\alpha_i \sim Normal(\mu_\alpha, \tau_\alpha) \ \forall \ i \in 1, ..., 66 \tag{6}$$

Similarly, we believe that time-specific treatment effects may be unique to each period but similar over time. We operationalize this concept by modeling all time-specific treatment effects γ as coming from a common distribution, with *hyperparameters* μ_γ and τ_γ.

$$\gamma_t \sim Normal(\mu_\gamma, \tau_\gamma) \ \forall \ t \in 0, 4, 8, 12, 24, 36 \tag{7}$$

μ_γ is the *grand mean*, the overall treatment effect; τ_γ is the variation in treatment effects over time; and each individual γ_t is a time-period specific treatment effect. This approach has a natural smoothing effect: any extreme estimates of γ_t will be partially-pooled back toward the grand mean μ_γ.

We assign the following prior and hyperprior distributions:

$$\mu_\alpha \sim Normal(5, 5) \tag{8}$$

$$\tau_\alpha \sim Cauchy^+(0, 30) \tag{9}$$

$$\mu_\gamma \sim Normal(0, 5) \tag{10}$$

$$\tau_\gamma \sim Cauchy^+(0, 30) \tag{11}$$

$$\theta \sim Normal(0, 1) \tag{12}$$

The normal distributions around the individual and treatment effects allow us to guide the model to the appropriate range of parameter values, but with wide enough variance (5 in each case) to let the model find its own way in that range. Half-Cauchy priors on the variance parameters are weakly informative, with much of their mass around zero but gentle slopes in their tails, which have been shown to be effective prior distributions for variance parameters (Gelman, 2006).

3.5 Model estimation and results

We estimate this model with *Hamiltonian Monte-Carlo* in Stan. Model code is appended to this document. This is a particular algorithm from a larger class of Markov Chain Monte Carlo algorithms, for more examples see Gelman et al. (2014).

Model results are displayed in Table 2. The table displays the mean of each of the parameter distributions along with the 50% posterior interval. Results suggest that using the Noom Monitor smartphone application during CBT-GSH may slightly decrease OBEs. There is evidence that the treatment effect varies over time, with the Noom effect being slightly more pronounced during stages 4, 8, 12 and 24 of therapy but decreasing by week 36.

	mean	25%	50%	75%
γ_0	0.18	-0.45	0.15	0.78
γ_4	-0.43	-1.05	-0.46	0.16
γ_8	-0.70	-1.33	-0.71	-0.10
γ_{12}	-0.65	-1.28	-0.68	-0.04
γ_{24}	-0.72	-1.34	-0.75	-0.11
γ_{36}	0.21	-0.42	0.19	0.82
μ_γ	-0.34	-0.98	-0.36	0.26
τ_γ	0.64	0.43	0.56	0.77

Table 2: *Table displays model results for Noom effects in all six time periods and grand mean and variance parameters.*

3.6 Model checking, comparison, and expansion

Before using our model to make inferences about time-specific treatment effects, we check its fit by comparing model-simulated OBE to data OBE. If model simulations do not track the data well, we may want to revisit our model's assumptions before trusting its inferences. If the model's simulations recover patterns in the data, we are more inclined to trust it.

Figure 3 displays OBEs in each period for each individual in each treatment group, from raw data (upper plots) and model simulations (lower plots). Black lines display means for each period. This suggests that the model is able to pick up on the key variables that determine OBE over time for the duration of this experiment.

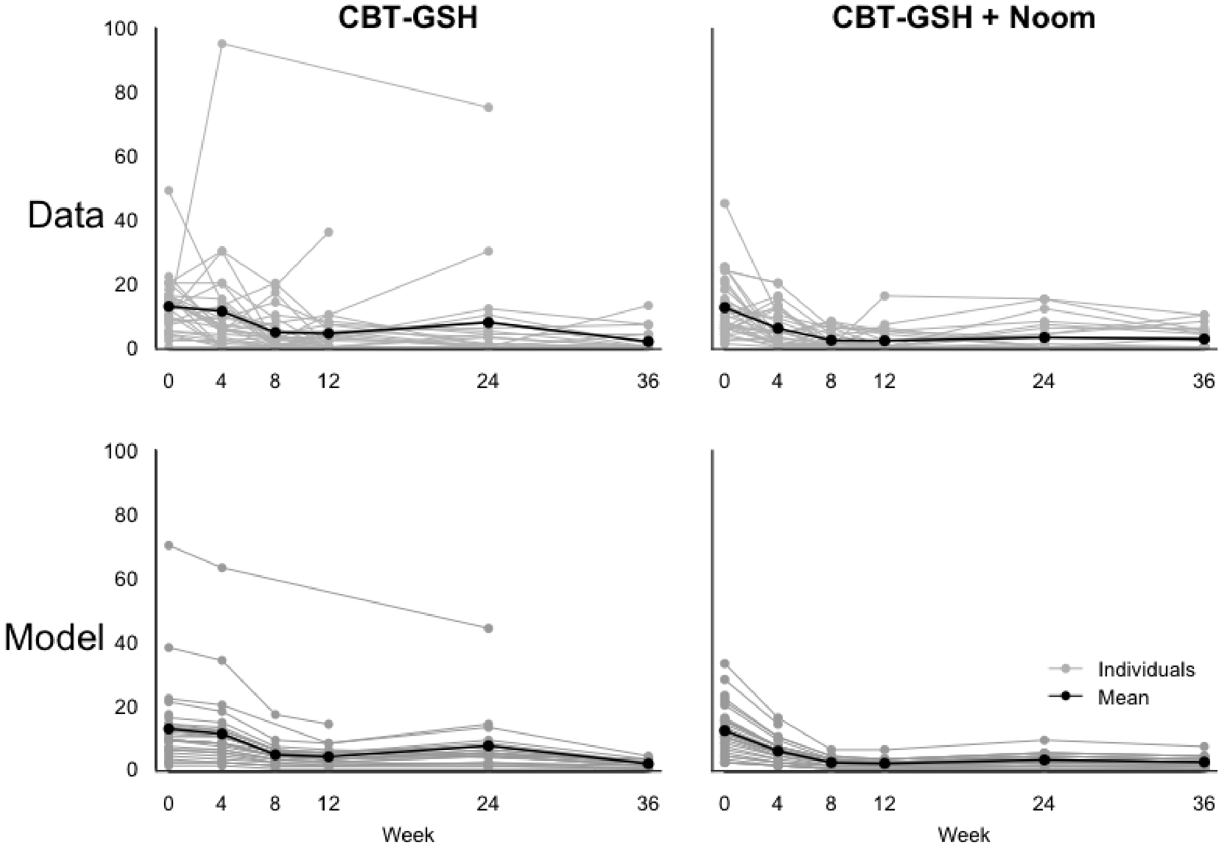

Figure 3: *This figure compares the original data in gray to data simulated from the model in black for the control and treatment condition. Fig.3 (a) and (b) in the upper row show OBEs in each period for each individual in CBT - GSH and CBT-GSH + Noom respectively. Black dots represent mean estimate for each period. Fig.3 (c) and (d) in the lower row show the simulated OBEs for each individual in CBT - GSH and CBT-GSH + Noom respectively. Black dots represent means from the simulated data for each period. The horizontal axis shows the week and the vertical axis shows the instances of OBE.*

Another way to check the fit of the model is by comparing simulated data directly against the raw data. This is called posterior predictive checking and is our preferred form of model checking. Figure 4 shows this for both treatment conditions. Simulated data for the Noom condition appear to better track the raw data than simulated data for the no Noom condition. This is unsurprising, since the no Noom condition tended to have more outliers, which we would not expect (or want) our model to pick up perfectly from such a small sample. If our model worked “perfectly” all the points would cluster along the sloping 45 degree line.

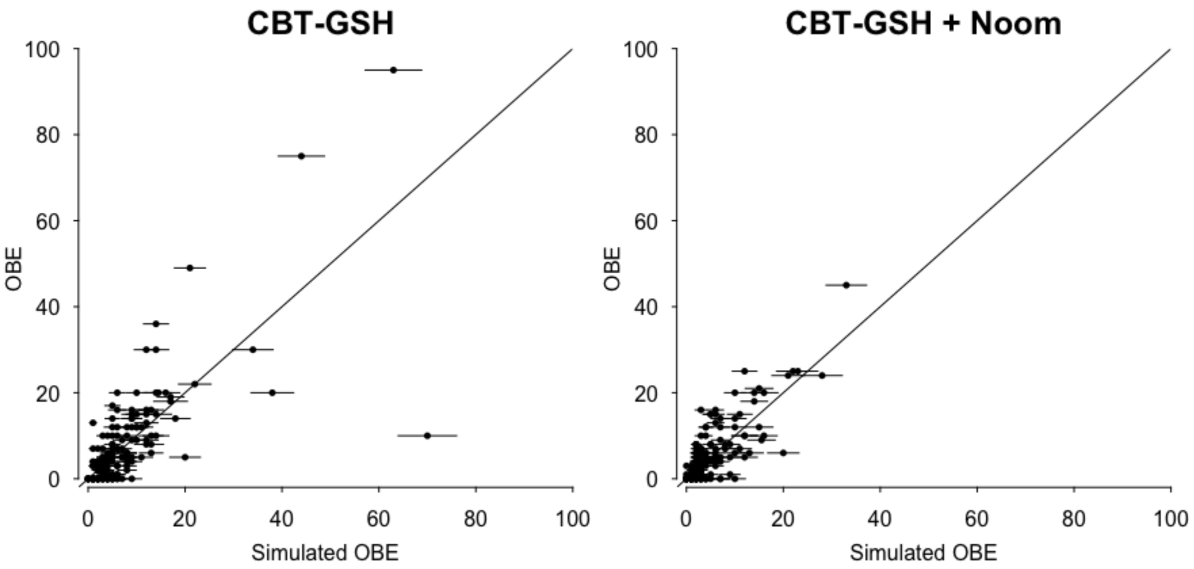

Figure 4: *This figure shows the posterior predictive checks by examining how well the model is able to predict the data. Fig. 4(a) and (b) show the predicted OBE vs. the actual OBE data for CBT - GSH and CBT-GSH + Noom respectively. The horizontal axis shows the simulated OBE and the vertical axis shows the actual OBE. The black dots are the points that represent this and the lines around them show 50% intervals around the predictions. The upward sloping line is the 45 degree line.*

We compare the distributions of OBE for each condition in each time period by plotting density curves over the histograms in figure 2, displayed in figure 5. Figure 5 shows that our model is able to broadly pick up on the patterns in the data over time and between treatment conditions. We see that the density curves clearly peak at lower values close to zero and have long tails, correctly capturing the pattern in the OBE data. We recover the pattern of excess zeros and the skewness of the data as well as the long tails of the OBE distribution in each time period.

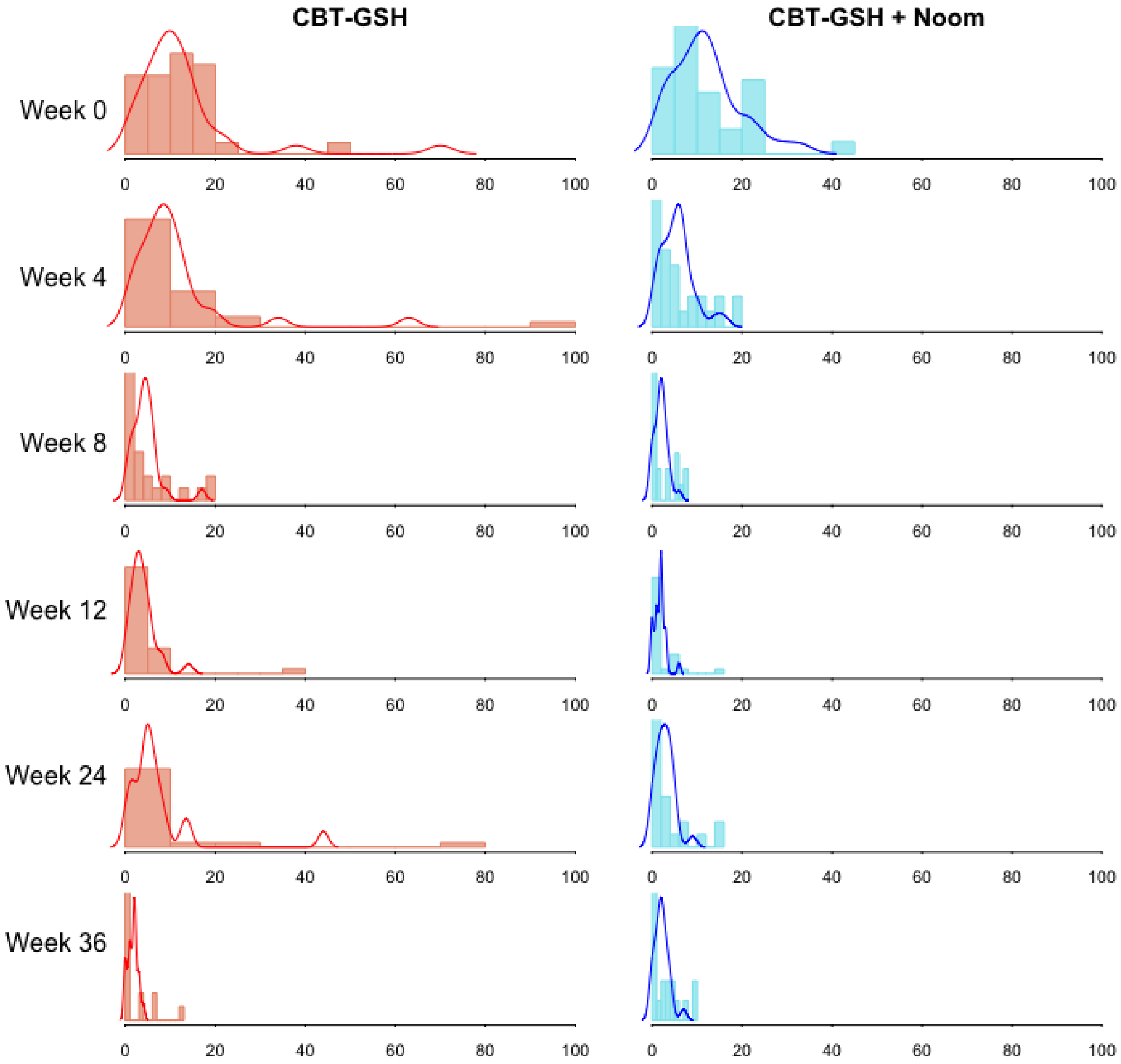

Figure 5: *This figure shows the density distribution of the predicted values overlayed on the histograms from figure 2. The histograms display the distribution of OBEs in each condition in each stage of the treatment. The orange and light blue histograms show the distribution of OBEs for the CBT-GSH and CBT-GSH + Noom group respectively. The red and blue lines show the predicted density curves obtained from the simulations.*

Figure 6 displays the simulated OBE for both treatment groups (upper plot) and smoothed treatment effects (lower plot). In each measurement period, simulated OBE are higher for the CBT - GSH condition than for the CBT - GSH + Noom condition, with some of the difference likely attributable to use of the Noom Monitor smartphone app. This shows that the app has an effect on lowering episodes of binge-eating.

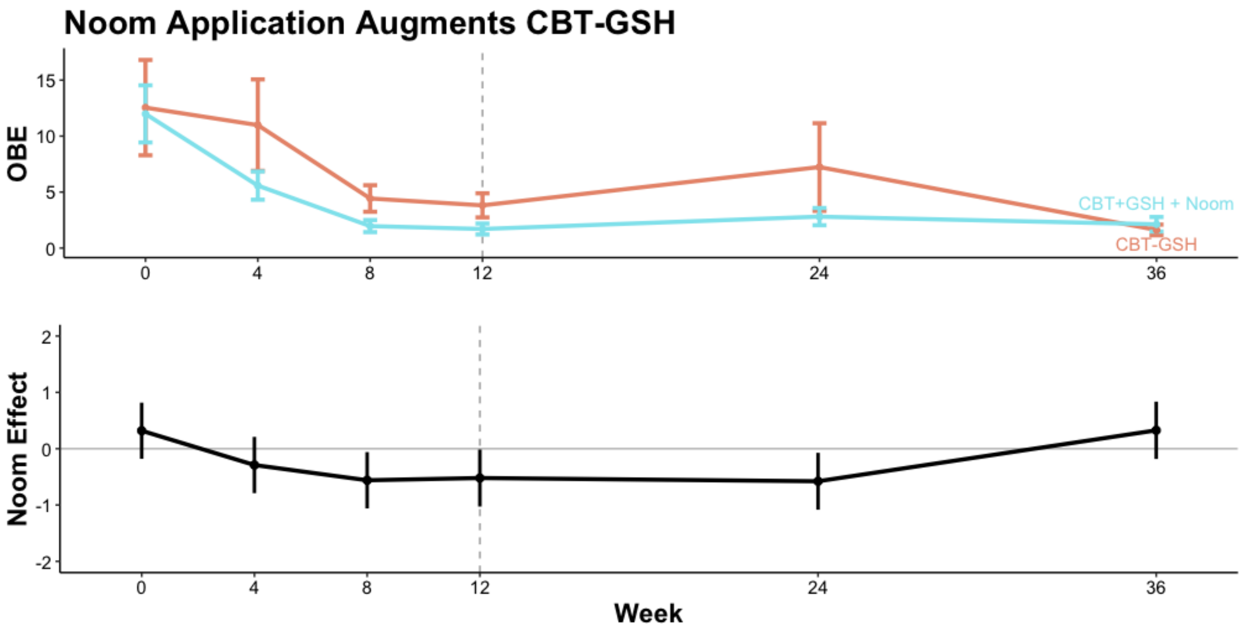

Figure 6: *This figure summarizes the treatment effects and the OBE patterns across the time period under consideration. Fig. 6(a) the upper plot shows the simulated OBE in each time period. The CBT - GSH condition is in orange and the CBT - GSH + Noom condition is in orange. The bars around each point show the 95% interval. Fig. 6(b) the lower plot shows the treatment effect for each period with the bars showing the 50% interval. The horizontal axis shows the time period and the vertical axis for Fig.6 (a) shows the instances of OBE and for Fig.6 (b) shows the treatment effect of the Noom app.*

4 Discussion

In this section we discuss some of the major benefits of using a Bayesian approach to the analysis of Randomized Controlled Trials.

4.1 Heterogenous Treatment Effects

Among the advantages of Bayesian modeling is its ability to capture heterogeneous treatment effects. For example in our study, we control for variation in demographic variables and variation in the treatment across the time. This approach is commonly known as a varying slope - varying intercept model (Gelman & Hill, 2006). Our approach of hierarchical modeling and partial pooling (Gelman et al., 2014) allows us to accomplish this naturally.

Hierarchical modeling has two closely related meanings (Feller & Gelman, 2015). Hierarchies can explain a hierarchical data structure like spatial or temporal variation. In the RCT we consider in this paper, each treatment period is treated as a level and so we obtain different treatment effects for each period. Hierarchies also describe how parameters are modeled. We model our treatment effects for each level or category in the data, to come from a common underlying prior distribution (Gelman et al., 2014). This is the idea of *partially pooled* estimates of treatment effects. Partial pooling in a Bayesian context is different than traditional complete pooling or no pooling approaches.

In non-Bayesian methods, researchers have two options: complete pooling and no pooling. With complete pooling, the categories in data are completely interchangeable, thus ignoring the uniqueness of categories. With no pooling, each category is treated as independent from the others, ignoring the interrelatedness of each of the categories. For example, if an RCT was conducted in multiple locations by different staff, the researcher must assume that the treatments are entirely identical in a complete pooling model, ignoring the differences in execution that may have taken place by differing staff. In the no pooling model, the RCT would assume that each treatment would be entirely different across locations, despite having close similarities in how the treatments were applied. Setting up a prior distribution allows researchers to solve this problem by saying that the treatment effects across the locations have a common mean, but vary based on location. Thus partial pooling is often a better representation of data as it takes into account both the uniqueness and interrelatedness of categories within a hierarchical model.

4.2 Making uncertainty explicit

Modeling uncertainty in Bayesian statistics is fundamentally different from other non-Bayesian methods (Classical and frequentist statistics). In frequentist statistics, uncertainty in parameter estimates is commonly represented as a 95% confidence interval (CI). Under the assumption of a very large number of hypothetical replications of the data, 95% of the estimated parameter values are expected to fall within the confidence interval. Additionally the frequentist confidence interval relies on many other assumptions and can be misinterpreted (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). In Bayesian statistics, posterior distributions of the parameters are derived by multiplying prior and likelihood distributions. Thus these parameters have full distributions with probabilistic uncertainty, instead of point estimates with confidence intervals. The Bayesian approach does not rely on asymptotics or assumptions about the distribution of error terms; it models uncertainty explicitly.

5 Conclusion

Bayesian Data Analysis is a powerful tool for incorporating data and prior information into flexible statistical models. It is well-suited to analyze RCTs, where effects are small, prior information is plentiful, and data may have hierarchical structures. In this paper, we have explained the steps of Bayesian Data Analysis and shown how they can be used to analyze an RCT that evaluates treatments for binge-eating disorder (BED) and bulimia nervosa (BN). We argue that Bayesian methodologies are well suited to analyzing RCTs of eating disorder behavior because they allow researchers to model uncertainty and heterogeneous treatment effects explicitly. We demonstrate our approach by analyzing the impact of a smartphone app on binge-eating behavior by fitting a hierarchical Poisson model with individual-level and time-level effects, and time-varying treatment effects. These unique representations of assumptions can be useful to empirical researchers in general and psychiatrists and clinical psychologists in particular.

6 Appendix

Example priors

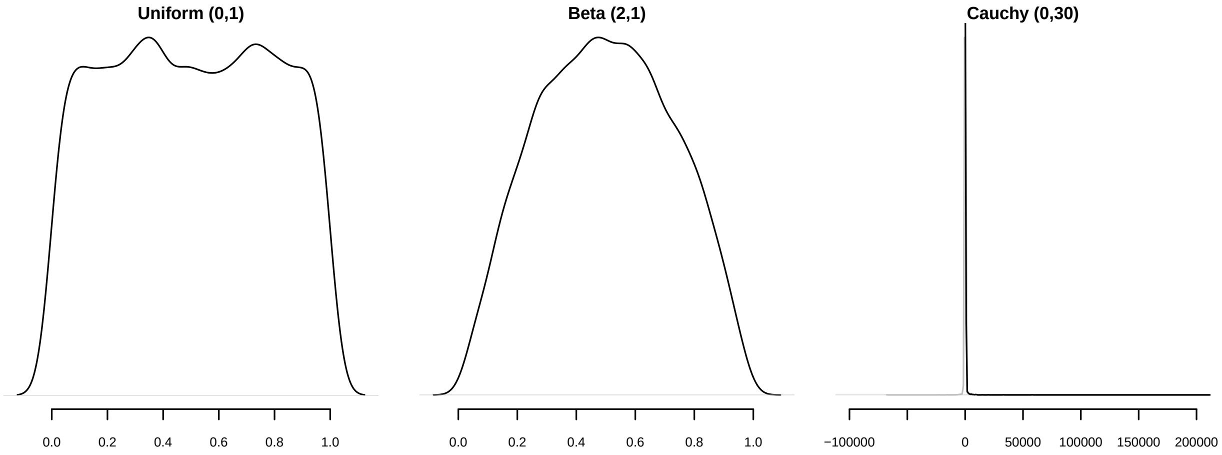

Figure 7: *This figure shows some examples of common prior distributions. Fig. 7(a) shows a Uniform (0,1) distribution. Fig. 7(b) shows a Beta (2,1) distribution. Fig. 7(c) shows a Cauchy distribution, where the black line shows the positive half and the gray line shows the negative half.*

In Figure 7, we show some sample prior distributions. The Uniform (0,1) distribution (Fig. 7(a)) can be used when we know the quantity of interest is constrained to be between 0 and 1 but believe that all values between 0 and 1 are equally likely. If our quantity of interest is between 0 and 1 but unlikely to take extreme values we can use a Beta (2,1) (Fig. 7(b)) prior as shown. Cauchy^{+}(0,30) is the half-Cauchy distribution with location 0 and scale 30 (Fig. 7(c)) . This is a good prior for variances because along with restricting the distribution to the positive real line it places most of the mass at 0 but allows for long smooth tails that the HMC algorithm can explore.

References

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Betancourt, M. (2018). Calibrating model-based inferences and decisions. *arXiv preprint arXiv:1803.08393*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Feller, A., & Gelman, A. (2015). Hierarchical models for causal effects. *Emerging Trends in the Social and Behavioral Sciences: An interdisciplinary, searchable, and linkable resource*.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2017). Visualization in bayesian workflow. *arXiv preprint arXiv:1709.01449*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, *1*(3), 515–534.

Gelman, A. (2013). Commentary: P values and statistical practice. *Epidemiology*, *24*(1), 69–72.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). CRC press Boca Raton, FL.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Grotzinger, A., Hildebrandt, T., & Yu, J. (2015). The benefits of using semi-continuous and continuous models to analyze binge eating data: A monte carlo investigation. *International Journal of Eating Disorders*, *48*(6), 746–758.

Hildebrandt, T., Michaelides, A., Mackinnon, D., Greif, R., DeBar, L., & Sysko, R. (2017). Randomized controlled trial comparing smartphone assisted versus traditional guided self-help for adults with binge eating. *International Journal of Eating Disorders*, *50*(11), 1313–1322.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*(5), 1157–1164.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan.* Academic Press.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, *23*(1), 103–123.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Stan Development Team. (2015). *Prior choice recommendations.* `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`. (Accessed: 2018-04-27)

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

2.17.3)

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2). Reading, Mass.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.