[redacted]

Evaluating the utility of Bayesian BART models for hierarchical parameter estimation in the Columbia Card Task

[redacted]

School of Communication and Culture, Aarhus University

MSc Cognitive Science

Decision Making Exam, 13/01-2023

Abstract

(ESH, NMA & LT) This paper examines the utility of two different Bayesian Balloon Analogue Risk Task (BART) models for estimating group-level parameters related to risk-taking behaviour in the hot version of the Columbia Card Task (CCT). With a point of departure in a CCT experiment involving crack users and control participants, we aim to examine whether estimates of risk propensity and behavioural consistency, as measured with the CCT, can be reliably approximated using models designed for a different, but theoretically similar task (BART). This is achieved by comparing the abilities of the two models to accurately estimate group-level differences, which entails fitting the models to the data, and assessing their efficacy in terms of parameter recovery and predictive accuracy through posterior predictive checks. The results of these analyses are used to evaluate the generalizability of the models and discuss the merits and limitations of applying them to CCT data. Based on the results we suggest that, with slight modifications, these models provide an adequate parameter estimation of CCT data. However, we find that this adaptation has several flaws. One such flaw is subpar predictive accuracies, which highlight the need for further improvements and modifications to the models in question, if they are to be applied to CCT data.

1.0 Introduction

(ESH) A central challenge for all living organisms is to make the decision with the best payoff when facing uncertain and risky situations. Is it the better decision for a thirsty antelope to approach a river full of crocodiles, or take the chance that another water resource will show up in time? Likewise, is the most optimal decision for a house owner to sell at a time where market prices are upward-trending, or wait in the hope that they will increase even more? The antelope might get attacked by crocodiles, or it might die of thirst; the house owner might sell for a price much lower than what she could have got, or risk a market crash. In both scenarios, the preference for either acting or waiting is related to the decision maker's risk propensity, i.e. how willing one is to accept uncertainty and potential loss in relation to an expected outcome (Josef et al., 2016). Quantifying risk propensity is tricky, as it is implicitly integrated in the decision making process and is intertwined with other external- and internal factors.

(NMA) There is currently no general consensus as to how risk propensity should be measured, and different approaches appear to reach different conclusions (Pedroni et al., 2017). In the current paper, we apply two methodologically different modelling approaches to estimate parameters of risk propensity and behavioural consistency in a cohort of crack cocaine users as opposed to one of controls, as measured by the experimental "hot" Columbia Card Task (CCT). We do this with the aim of investigating the applicability and generalisability of different methods for analysing data from different risk measures applied in cognitive science. This way, we hope to be able to make conclusions regarding the way latent concepts such as risk propensity can be measured and discussed in the most satisfactory manner.

1.1 Risk propensity in decision making

(LT) Decision making under risk demands specific decision making strategies (O'Doherty et al., 2017). In this paper, we define risk in decision making as the perceived level of consequences and uncertainty combined (Aven, 2012). Classical approaches to decision making under risk such as expected utility theory assume that decision-makers are rational agents who always choose the

objectively optimal decision. In other words, people are expected to weigh evidence logically and choose the option with the highest expected utility no matter the risk (Machina, 1987). However, the reality is that people often tend to be overly cautious when faced with the possibility of loss, especially as stakes increase (Ert & Yechiam, 2010). This phenomenon is known as loss aversion, which occurs when the prospect of loss has a larger impact on choice than an equivalent prospect of gain (Kahneman & Tversky, 1982).

(ESH) Similarly, people tend to choose safe choice options rather than uncertain ones, even when the expected utility is higher for the uncertain outcome, a phenomenon known as risk aversion (Klaus et al., 2020). Loss- and risk aversion has had an important impact upon decision making research and increased the incentive to reach an understanding of how risks influence choice (Nicholson et al., 2001). In the current paper, we are interested in how latent measures of risk propensity can be estimated from behavioural experiments. In that regard, it is important to keep in mind that whilst the concepts are distinct, risk propensity is closely related to loss aversion and risk aversion, and experimental isolation can be difficult.

1.1.1 Factors affecting risk propensity

(NMA) In contrast to the general predictions by theories on decision making under risk, observations from real life show how some people are willing to run risks that others avoid. Whilst e.g. prospect theory generally predicts that the risk propensity adopted by an individual is situational (Nicholson et al., 2001), research also suggests that individual people have relatively stable "general" risk propensity profiles (Frey et al., 2017; Josef et al., 2016). Factors which are thought to contribute to such a general risk propensity profile include e.g. gender and personality. For instance, some correlations has been found between risk propensity measures and character traits such as openness to experience (Highhouse et al., 2022), extroversion (Nicholson et al., 2001), and impulsivity (Nicholson et al., 2001; Lauriola et al., 2014; Penolazzi et al., 2012). Specifically, risk propensity seems to correlate strongly with sensation-seeking behaviour (which by some is also considered a personality trait) (Lauriola et al., 2014; Penolazzi et al., 2012). In the current design, we are interested in whether we can detect different levels of risk propensity in chronic crack cocaine users versus controls.

(LT) In this context, we expect drug addicts to portray higher levels of risk propensity, due their specific sensation-seeking, impulsive personality profiles (Lauriola et al., 2014; Nemeth, 2009), and because drug dependency and consumption correlates with high-risk behaviour (Kluwe-Schiavon et al., 2016; Wittwer et al., 2016; Koffarnus & Kaplan, 2018). Relatedly, it is relevant to mention the second measure we estimate in the current study: behavioural consistency. Whilst the phrase sometimes is used in relation to the consistency of risk propensity levels between different tasks and/or domains (e.g. Josef et al., 2016), it is here understood as the within-task consistency of risk-propensity related behaviour. For different tasks, the level of behavioural consistency is suggested to indicate how deliberate (high consistency) or hazardous (low consistency) an approach participants choose; or to indicate how explorative (low consistency) or exploitative (high consistency) a strategy is applied (Bishara et al., 2009). In drug addicts, behavioural consistency is assumed to be lower than for controls, as they are supposed to be less inclined to perform deliberate evidence weighting and be more inclined to guide their choices according to immediate reward and punishment (Bishara et al., 2009; Kluwe-Schiavon et al., 2016).

1.1.2 How risk propensity is measured

(ESH) As stated, risk propensity is an implicit psychological construct in a choice situation. Thus, it is difficult to measure directly. In the literature, two main approaches to measuring risk propensity are applied: Rating scales and experimental laboratory tasks (Figner et al., 2009). Rating scales include for example the risk propensity scale (RPS) and the domain-specific risk taking scale (DOSPERT) (Meertens & Lion, 2008; Blais & Weber, 2006). A shared characteristic of rating scales is that they rely on self-report, which generally suffer from validity issues and are often only weakly correlated with corresponding behavioural measures (Dang et al., 2020). Experimental paradigms, on the other hand, rely on actual behavioural measures rather than introspection and are especially relevant in economic- and psychological research. Some examples of these are simple economic games such as monetary lotteries (e.g. von Helversen & Rieskamp, 2013) and sequential paradigms such as the Balloon Analogue Risk Task (BART) (Lejuez et al., 2002) and the Columbia Card Task (CCT) (Figner et al., 2009).

(NMA) Experimental paradigms are generally considered more reliable than self-report measures (Dang et al., 2020), but struggle with being highly artificial (Holleman et al., 2020). An important issue in both self-report and behavioural measures of risk propensity is that they do not necessarily measure what researchers think they measure. For example, risk propensity might be domain-specific, meaning that it varies for different areas of life, and possibly with different domain-profiles for different groups (Weber et al., 2002; Hanoch et al., 2006). A related finding is that even slight differences in design across experimental paradigms might motivate different decision making strategies and, following this, different risk propensity profiles. For instance, six different experimental paradigms designed to assess risk propensity (BART, decisions from experience, CCT, adaptive lotteries, marbles task, MPL and Holt and Laury gambles) did not consistently capture risk preference in the same experimental groups (Pedroni et al., 2017). These many variations could suggest that various methods for measuring risk capture fleeting states of risk propensity rather than stable tendencies (Frey et al., 2017). As such, developing methods for analysing data on risk is no easy task. In the present article, we are interested in investigating whether modelling approaches developed to measure risk in the BART can be used to measure risk in the similarly structured hot CCT.

1.1.3 The Balloon Analogue Risk Task (BART) and the Columbia Card Task (CCT)

(LT) The two experimental tasks relevant for this paper are the Balloon Analogue Risk Task (BART) (Lejuez et al., 2002) and the Columbia Card Task (CCT) (Figner et al., 2009). Both are sequential, computerised tasks used for assessing risk propensity in various populations. The oldest of the two tasks, namely the BART, is a task where participants act under risk in the sense that they have to balance the benefits of monetary gains against potential losses (Lejuez et al., 2002). The main objective of the task is to maximise monetary gains by inflating a virtual balloon (Lauriola et al., 2014). Every time the balloon is pumped, a certain value is added to a participant's temporary bank. The money is permanently saved only if the participant stops the trial and "banks" their earnings before the balloon bursts. If a participant pumps until the burst point, all earnings gained on that trial are lost.

(ESH) The point at which the balloon bursts differs between trials, and so it is not possible to learn the burst patterns in the experimental design. As such, all pumps in the experiment are

associated with both increased risk and increased gains. In the original version of the task, the probability of the balloon bursting on the first pump on a trial was either ⅛, 1/32, or 1/128. The probability of reaching the burst point increased linearly thereafter, given that the first pump did not lead to a burst. All participants went through a total of 30 balloons. Usually, the average number of pumps on non-burst trials, known as the average adjusted pumps, is used as the measure of participants' risk propensity. Differences in this number are thought to reflect individual differences in risk propensity.

(NMA) The CCT is similar to the BART in the sense that it also has a dynamic probability structure of risk, and in that it includes the option to stop a trial when continuing is deemed too risky (Figner et al., 2009). The general task design involves an array of 32 downwards facing cards. Most of these are gain cards, while either one or three are "bad" cards, depending on the condition of the trial. On each trial, the participant is given information regarding the gain value of the good cards, the loss value of the bad cards, as well as the amount of loss cards in the array on a given trial. The objective of the task is for the participant to maximise monetary gains by flipping around as many good cards as possible. If a bad card is encountered, the loss number will be subtracted from the total gains so far in the game. In the hot version of the CCT, participants receive immediate feedback regarding what type of card they have flipped around. (LT) Thus, the increasing probability of encountering a bad card is directly inferrable in this version of the game. The trial is forcibly ended if a bad card is encountered. Due to the element of immediate feedback, the hot version of the CCT is associated with affective rather than deliberate processing of risk. As the probability of encountering a loss card increases as more cards are flipped, the mean number of flipped cards across trials was originally used as the measure of risk propensity.

(ESH) Since they were first developed, both the BART and the CCT have been widely applied as experimental methods to measure risk propensity. As pointed out in a meta-analysis by Lauriola et al. (2014), correlations between the adjusted pump measure in the BART-task and real life risk behaviour such as drug abuse and gambling have been found. Also, group comparison studies have found that traditionally risk seeking groups, such as cocaine addicts, display a higher number of adjusted pumps than control groups, which is argued to be proof of the validity of the

BART. The CCT has also been subject to assessment of validity and relevance for application within research on risk propensity. In a study investigating the test-retest reliability in commonly applied risk propensity tasks, including the BART, (Buelow & Barnhart, 2018), fewer practice effects and greater reliability was found for the CCT than the more commonly applied Iowa Gambling Task (IGT) (Bechara et al., 1994). The CCT was also suggested as a more stable method for assessing decision making processes than the other tasks included in the study. Despite these findings, some inherent issues of the task designs are likely to reduce their validity. In this paper, the one issue that will be pointed out is the aspect of forcibly ended trials in both experimental tasks.

(NMA) The key issue with forced termination of trials on the BART and the CCT is connected to the way risk propensity is measured on both tasks. As pointed out in a paper by Coon & Lee (2022), using the average number of pumps/flipped cards as the representation of risk propensity complicates all efforts of making inferences regarding risk propensity. Once a burst or a loss card forces the participant to stop making choices in the game, we cannot know the true point at which the participant would have deemed the game too risky to continue. As such, their true risk propensity is concealed. As a matter of fact, if not accounted for, this issue likely leads to underestimation of an individual's risk propensity, as well as improper judgement of their behavioural consistency (Coon & Lee, 2022; Huang et al., 2013; Pleskac et al., 2008). The adjusted pump measure for the BART is one of several attempts to overcome this issue. However, although excluding all trials that ended in a burst seemingly removes the problem of concealed risk propensity on certain trials, this method greatly reduces the statistical power of all analyses attempting to make inferences regarding individuals' risk propensity.

(LT) Additionally, if a great proportion of trials must be discarded for an experimental task to be reliable, the validity of the task should be questioned. For the CCT, a proposed solution to the issue was to create a different version of the task, called the "warm" delayed feedback version (Huang et al., 2013). In this version of the task, feedback about the nature of the chosen amount of cards on a trial is given after the participant has decided upon the number of cards they wish to flip. As such, it is virtually possible to choose more cards even after the loss card was chosen. A similar solution for the BART is called the automatic version of the BART (Pleskac et al.,

2008). Here, participants decide at the very start of a trial how many pumps they wish to make, and subsequently observe the sequence of pumps unfold. These alternative versions thus expose participants' true intended number of pumps. Yet, they are problematic, as such structural alterations to the experimental setup are likely to alter the cognitive processes involved in the task (Wright & Rakow, 2017). The problems pointed to in this section highlight the need for developing theoretically sound ways of analysing data from the two experimental tasks presented, namely from BART, and from the hot version of the CCT. The two alternative methods presented and discussed in this paper are built on frameworks of cognitive and statistical modelling of cognitive processes.

1.1.4 Cognitive versus statistical modelling of cognitive processes

(ESH) As with the variety of measures that exist for assessing risk propensity, several ways of analysing data from experimental paradigms exist. Two of these include cognitive and statistical modelling. Despite serving complementary purposes within cognitive science, important distinctions between the two have to be made. Starting out with cognitive models, these are mathematical formulations of theories of cognition (Busemeyer & Diederich, 2010). They are directly derived from principles of cognition, which sets them apart from generic statistical approaches. The aim of a cognitive model is to infer certain latent cognitive constructs from behavioural data and thus describe human behaviour and cognition on a more abstract level than standard statistical approaches. Cognitive models can be used for making predictions about human behaviour that goes beyond observed data. On the one hand, the mathematical characters of these models have been used as an argument for their logical validity and generalisability. On the other hand, cognitive models are based on assumptions regarding the cognitive processes involved in various tasks, which has been pointed to as a major drawback of the models (Coon & Lee, 2022).

(NMA) Yet, it should be mentioned that cognitive models simply seek to provide the best representation of a given cognitive process, and that their estimates should always be considered an approximation (Busemeyer & Diederich, 2010). In contrast, the statistical approach aims to describe and quantify the observed data, and not to directly validate theories of cognition (Kennedy et al., 2019). Unlike cognitive models, the statistical approach does not involve the

commitment to theoretical assumptions regarding cognition. Inferences and conclusions on cognition are done by evaluating cognitive theories against the data. Despite the two approaches having fundamentally different immediate goals, both seek to make conclusions regarding cognitive processes. Ideally, both models should therefore produce complementary findings when used to analyse a certain cognitive process.

1.1.5 Motivation for modelling data from hot CCT with BART-models

(LT) In the current paper, we compare risk propensity and behavioural consistency profiles between two distinct cohorts through data obtained with the hot CCT. As previously stated, our central interest is to assess whether models initially designed for modelling data from BART can generalise to this type of data. This is motivated by several considerations. The most important motivation is that similarly to BART, hot CCT struggles with the issue of forced terminations skewing the risk propensity and behavioural consistency estimates. To our knowledge, no modelling frameworks which sufficiently accounts for this issue have been developed for the hot CCT. As the hot CCT and BART share central characteristics, we hope that novel modelling techniques which have been applied to deal with the forced termination issue in BART might be generalisable to the hot CCT. The primary argument for considering BART and hot CCT structurally similar is that the relative risk of forced termination increases for each increment (pump or card turn), simultaneously with a decrease of the relative gain (Buelow & Blaine, 2015). Further, both tasks include immediate feedback, thus supposedly involving the affective systems.

(ESH) Taken together, this might suggest that BART and hot CCT activate similar decision making strategies and risk propensity profiles (Buelow & Blaine, 2015). However, their comparability in relation to how risk propensity is evaluated is questionable. Recent research suggests that the risk propensity measures obtained on BART and hot CCT are quite different (Pedroni et al., 2017). A further motivation for the current modelling approach is thus that if data from hot CCT can be modelled with a similar modelling technique as BART data, this could serve two purposes: 1) To evaluate the validity of the BART models in their ability to assess reliable measures of risk propensity and behavioural consistency when modelling data that should be compatible, and 2) to evaluate whether applying models of a more similar structure on

data from different tasks could make the estimates of risk propensity and behavioural consistency more comparable. We are specifically applying two different models developed for BART data, the “cognitive BART model” and the “statistical BART model”, which approach the problem of modelling the latent factors of risk propensity and behavioural consistency in different manners.

1.2 Models employed for this study

1.2.1 The Cognitive BART model

(NMA) The first of the two models that will be applied and compared in the paper at hand is the two parameter BART model (henceforth referred to as the cognitive BART model), a cognitive model formulated by Wallsten et al. (2005) and further examined and discussed in a paper by van Ravenzwaaij et al. (2011). The cognitive BART model is a reduced version of Wallsten et al.’s (2005) preferred model out of ten cognitive models for the experimental task. The aim of Wallsten et al. (2005) was to provide a framework for investigating the cognitive processes underlying decision making under risk. The ten cognitive models of the BART presented in their paper differed in terms of the assumptions made in regards to the decision making processes involved in the task. All ten models included parameters that quantify risk taking, learning rate on the basis of experience, and behavioural consistency (van Ravenzwaaij et al., 2011). Three classes of models were created (Wallsten et al., 2005). One was a baseline class free of any psychological interpretations, which was intended to function as a statistical baseline against the other models. The last two classes were intended to be more plausible, seen from a cognitive viewpoint. The first of these two included so-called target models. These models assumed that decision makers learn on every choice in the BART. However, no deeper evaluation of the task was assumed to be kept throughout the game.

(LT) Also, these models made the assumption that decision makers held representations of an optimal target number of pumps prior to commencing the first trial, and that this target number subsequently influenced each decision maker’s probability of ending the trial throughout the game. The last class of models were categorised as learning and evaluation models. These models assumed individuals could learn and update their opinions towards the best number of balloon pumps between trials. All of the models were inspired by the Expectancy-Valence model

made for the Iowa Gambling task. The models were compared on the basis of the Akaike information criterion (AIC: Akaike, 1973) and on the correlation between the models' free parameters and various external measures of risk. The best model had four parameters and suggests that decision makers incorrectly do not update their view on burst probabilities throughout trials (van Ravenzwaaij et al., 2011). Furthermore, it assumes that they decide the number of pumps they will make before a trial begins, and that this number is not updated during the trial. Lastly, this model proposes that decision makers keep the same sensitivity to different outcome types throughout the game and revise probabilities in a Bayesian fashion. In their paper, van Ravenzwaaij et al. (2011) found indications of a need for simplifying the four-parameter model. This lead to the development of the cognitive BART model of risk.

(ESH) The cognitive BART model contains two free parameters, one governing risk propensity (γ) and the other behavioural consistency (β.) As with the four-parameter model of Wallsten et al. (2005), this version assumes a constant probability (p) of a burst throughout trials and that this probability is known to participants. The model includes two equations, one representing the number of pumps an individual considers to be optimal (ω) and the other the probability that they will continue the game at any given choice on a trial (θ). The first equation is formalised as such:

$$\omega \leftarrow \frac{-\gamma}{log(1-p)}$$

It follows from the equation that, the higher the risk propensity (γ) an individual has, the larger number of cards they will consider to be optimal (ω). In general, participants will consider a lower number of cards to be optimal if the probability of encountering a loss is high. This relationship is visualised in *Fig. 1*.

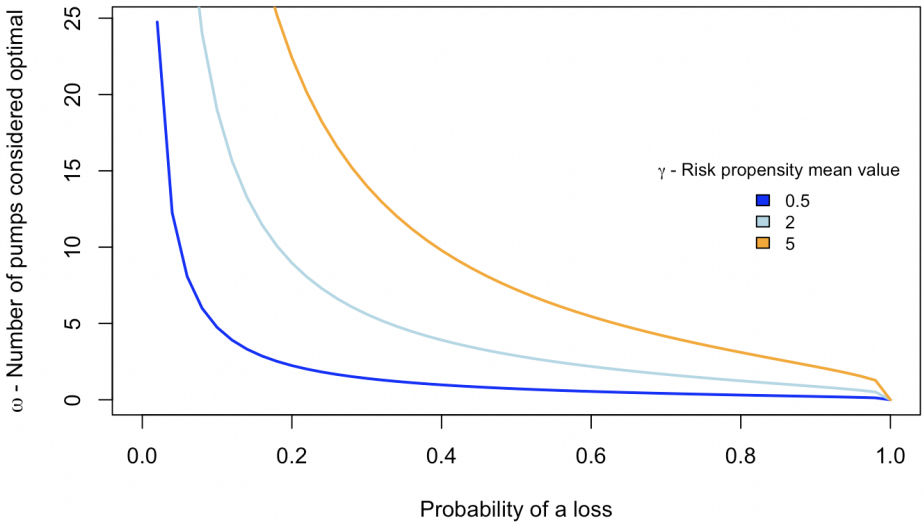


Figure 1: The figure illustrates the mechanisms underlying the equation for the number of pumps considered optimal. Generally, higher probabilities (*p*) of encountering a loss card leads to decreased numbers of cards considered optimal. Individuals with higher risk propensity will consider a higher number of cards to be optimal than individuals with lower risk propensity.

The second equation is formalised as such:

$$\theta_{jk} \leftarrow 1 - \frac{1}{1 + exp(\beta * \omega))}$$

(NMA) Whether or not a subject chooses to keep pumping the balloon on the *k*th opportunity on a given trial *(j)* depends on their behavioural consistency and the number of pumps they consider optimal. In general, the probability of continuing the game is reduced as participants approach their target number of pumps. A participant with very high behavioural consistency is thus likely to behave in a deterministic fashion, meaning that they have a high probability of pumping up until their target number of pumps. Similarly, their probability of pumping after their target number of pumps is virtually nonexistent. A participant with low behavioural consistency, on the other hand, will display more varying responses relative to their optimal pump number. The mechanisms behind this equation is displayed in *Fig. 2:*

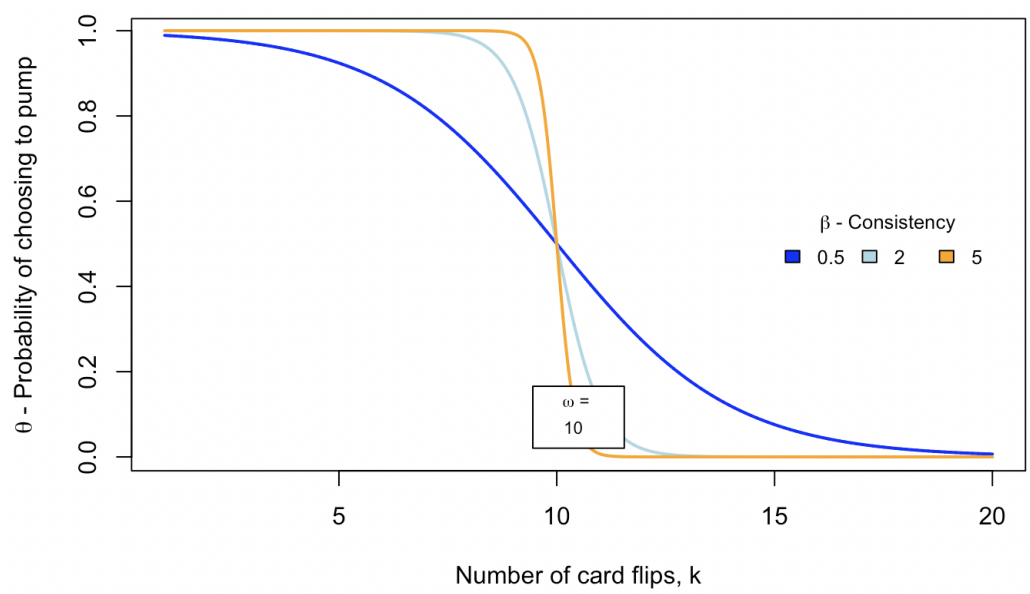


Figure 2: Individuals with lower behavioural consistency will display more variable responding relative to their optimal number of card flips. Individuals with higher behavioural consistency will act in a deterministic manner. This is visualised via the orange graph, which approximates a step function. In general, as people move closer to picking the optimal number cards, their probability of continuing the game becomes lower. The optimal number of card flips (ω) is set to 10 here for illustrative and comparative purposes.

Lastly, the observed decision (*d*) to pump the balloon on the kth opportunity within the jth trial is modelled as a Bernoulli trial:

$$d_{jk} \sim Bernoulli(\theta_{jk})$$

A graphical visualisation of the cognitive model can be seen in *Fig. 3:*

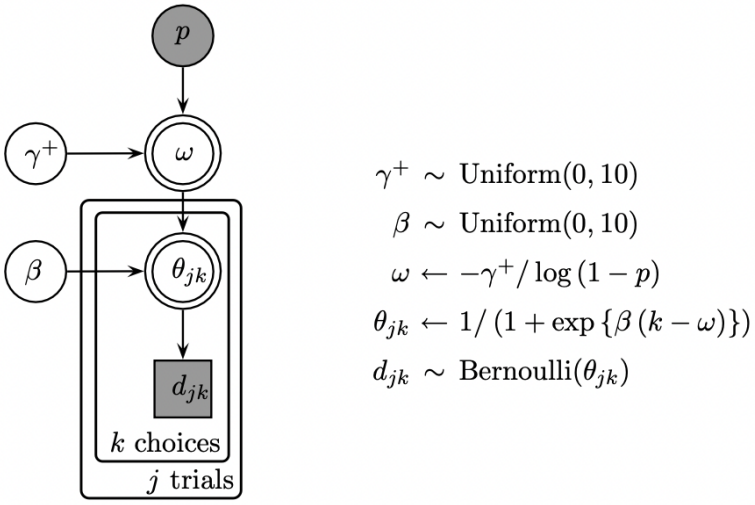


Figure 3: Note. Reprinted from "Bayesian Cognitive Modeling - A Practical Course" by Lee & Wagenmakers (2014.) Retrieved from https://webfiles.uci.edu/mdlee/LeeWagenmakers2013_Free.pdf. The plate notation illustrates the setup of the cognitive BART model presented by van Ravenzwaaij et al. (2011). In addition to the equations described in the text so far, the implemented priors for the risk propensity parameter (γ) and behavioural consistency (β) are shown. The shaded nodes represent observed values, while unshaded represent unobserved ones. Stochastic variables are represented by single-bordered nodes, while deterministic ones are double-bordered. Lastly, square nodes illustrate discrete variables, while circular ones represent continuous ones. As evident from this plate notation, only the probability of continuing the game (θ) is updated with every choice on each trial.

1.2.2 The statistical BART model

(LT) The next model we will use in this paper was developed by Coon & Lee (2022) as an alternative to cognitive models of the BART, and is called the censoring BART model (henceforth referred to as the statistical BART model). This statistical model takes on a Bayesian approach towards measuring risk propensity and behavioural consistency in individuals, and introduces the statistical method of censoring in order to overcome the issue of forcibly ended trials while simultaneously maintaining the original BART structure. Coon & Lee (2022) argued for the need of their model on the basis of issues with the adjusted pumps measure. Such issues include reduced statistical power of the task as a consequence of dropping all burst trials. Secondly, they provide an example illustrating how dropping trials from the experimental task is likely to lead to incorrect measures of risk propensity. The adjusted pump measure is insensitive to individual differences in risk and behavioural consistency where, due to the experimental task and choice of statistical analysis, the behavioural summaries seem identical. Also, they report, as

burst trials are likely to be reflecting higher risk since they also are likely to include a higher number of pumps, removing them artificially reduces the portrayed risk propensity of an individual. Other existing attempts of overcoming the problem are also deemed insufficient by the authors, leaving the censoring BART model.

(ESH) The aim of the censoring BART model is to statistically infer how many pumps an individual would have chosen had they not been forced to end the trial (Coon & Lee, 2022). This is done by measuring distributions of intended pumps for all individuals based on their behaviour on the BART. As such, this model includes all trials, except that burst trials are treated as censored measurements of how many pumps an individual would have chosen. The authors assume the distribution of intended pumps on the t^{th} trial (y'_t) to be sampled from a positively truncated Gaussian distribution. The mean of these distributions functions as the measure for individuals' risk propensity, while the deviance from the mean represents behavioural consistency. In the model specification, the mean of the intended pump distribution for the p^{th} participant is denoted as ϱ_p, and the standard deviation from the mean as β_p. Since the number of pumps has to be a non-negative integer, the sample from the truncated normal distribution is rounded to the nearest whole number, denoted with a $\cdot$ in the model formalisation. The positive Gaussian truncation is denoted as $Gaussian_+$. The formal definition of number of intended pumps (y'_t) is denoted as such:

$$y'_t \sim \lfloor \text{Gaussian}_+ \left(\rho_{p_t}, \frac{1}{\beta^2_{p_t}}\right) \rceil .$$

(NMA) Since the model is implemented in JAGS, the model specification includes precision instead of standard deviation. The precision is the reciprocal of the variance, and is considered standard practice when implementing models in JAGS. The latent construct of intended number of pumps (y'_t) is subsequently used to model the observed number of pumps in the data. The authors do this by right censoring all burst trials, so that the intended number of pumps on burst trials is higher than the observed number, and on bank trials, it is the same as the observed number of pumps. The point at which the balloon bursts is denoted as b_t, the observed number of pumps as y_t, and the entire censoring process is specified as such:

[redacted]

$$y_t = \begin{cases} y'_t & \text{if } y'_t < b_t \\ b_t & \text{if } y'_t \geq b_t. \end{cases}$$

Lastly, Coon & Lee (2022) set uninformative, flat priors for the risk propensity and behavioural consistency measures. This was done for the sake of simplicity in the presentation of the model. The priors were specified as:

$$\rho_i \sim Gaussian_+(0, \frac{1}{100^2})$$

$$\beta_i \sim Gaussian_+(0, \frac{1}{100^2})$$

(LT) The final model output includes the posterior predictive distributions for all participants, sampled from the inferred risk propensity of all individuals, as well as estimates for risk propensity and behavioural consistency. The posterior predictive distribution is a representation of the probability distributions of all numbers of pumps possible to make by an individual. These probabilities can thus be interpreted as inference of risk propensity. Parameter recovery for the model was good, and proved to be better than for the adjusted pump measure, where both risk propensity and behavioural consistency were poorly estimated. To be specific, risk propensity was often underestimated and behavioural consistency was overestimated for the adjusted pump measure. These findings are in line with the theoretical claims of the adjusted pump measure being an inadequate way of compensating for the forced ending of trials on the BART.

1.2.3 Theory behind the modelling framework

(ESH) The two models applied in this paper are implemented using the Gibbs sampling program JAGS (Plummer, 2003), which facilitates fully Bayesian inference by using computational sampling methods. It is handy when analysing models of high complexity, as it can handle a great variety of statistical distributions. JAGS takes a Bayesian model description as input, comprising model priors, likelihood, and model parameter relations. Subsequently, the method generates a sequence of random samples from the posterior distribution of the model parameters by using a Markov Chain Monte Carlo (MCMC) algorithm called Gibbs sampling (Ross, 2022).

These samples can then be used for estimating the posterior distribution of the data in question, and for making predictions based on these.

(NMA) The Gibbs sampling algorithm constructs a Markov chain, whose values converge towards a target distribution. This method is highly applicable when the joint distribution of some variables is unknown and difficult to sample from directly, but the conditional distributions of the variables in question are available and easier to sample from (Gelfand, 2000). A Markov chain is a stochastic model that describes a chain of events where the probability of each event in that sequence solely depends on the state in which a system was in the previous event (Agarwal, 2022). The creation of such a chain is accomplished by iteratively sampling from each of variables in a data set, given the current values of the others (Yildirim, 2012). Firstly, when running the sampling algorithm, values for all variables are initialised. This is often done in a random fashion. At each step in the iterative process, a new state is proposed by sampling from a pre-specified distribution.

(LT) The Gibbs sampling algorithm subsequently calculates the probability that the proposed new state will be accepted as the next state in the sampling process. If the proposed state is accepted, that state becomes the current state of the samples. The converged Markov chain would represent the distribution of the parameters we are trying to create a posterior estimate for. In this context, model convergence refers to the time point where the obtained samples approach the true underlying distribution of the variables in question. This usually happens as the number of iterations increases. As such, the samples produced become increasingly representative of the true distribution of the parameters. If a given model did not converge properly, the MCMC samples may not be representative of the true parameter distribution, and inferences based on the model estimates will most likely not be very reliable (Roy, 2019). Therefore, model convergence assessment is an important part of Bayesian inference in JAGS.

2.0 Methods

2.1 Data

(ESH) Access to the data applied in this paper was originally given in a previous collaboration between Liv Tollånes and Rodrigo Grassi-Oliveira, associate professor at the Translational Neuropsychiatry Unit (TNU) at Aarhus University. Written approval for extending the signed agreement to the current project, and for distributing the data amongst its collaborators, was provided from Rodrigo Grassi-Oliveira on request. The data used in this study stems from a Brazilian paper by Kluwe-Schiavon et al. (2016). This study used all three versions of the CCT to investigate how crack cocaine dependent women and healthy, teenage girls differed in terms of risk taking behaviour and in their use of information in risky decision making. Healthy adult women served as controls in the experiment. One of their findings specifically relevant for our study is that the presence of feedback reduced risk propensity in crack cocaine users to levels similar to those of the controls, who portrayed low risk through the conditions. This reduction in risk behaviour was argued to be a reflection of negative emotions associated with performance feedback, and not increased use and processing information available in the task.

(NMA) The participant group relevant to our study consisted of female crack cocaine addicts and a control group of healthy adult women between 18 and 50 years of age (males being excluded to avoid impact of gender differences). The participants in the addiction-group were diagnosed with crack cocaine dependence according to the Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition (American Psychiatric Association, 1994), had no psychotic symptoms, and no visual impairments. They were all recruited from a detoxification facility in southern Brazil. To avoid severe withdrawal symptoms, only those who had been a part of the detoxification programme for at least seven days prior to the study were eligible for participation. After excluding one participant due to moderate cognitive impairments, as measured according to the Wechsler Abbreviated Scale of Intelligence (Trentini et al., 2014; Weschler, 1999), the crack dependent group consisted of 27 women in total. Individuals in the control group had no history of substance abuse, and no visual disabilities. After removing one person due to a mood disorder, 20 healthy non-using adult women with regular incomes and

similar educational backgrounds were included in the study. Written consent for participating in the study was provided from all participants. Since the purpose of this paper is to compare the performance of two existing BART models on CCT data, minimal alterations were made to both models when fit to the hot CCT data. Detailed explanations of the modelling processes follow below. First, we explain how each of the two models were fitted to our data. Secondly, we explain the calculation of the methods on which we base the model comparison.

2.2 Hierarchical parameter estimation of CCT data with the cognitive BART model

(LT) We used the hierarchical extension of van Ravenzwaaij et al. 's (2011) cognitive BART model, as presented by Lee and Wagenmakers (2014). This version of the model included the addition of group level distributions for the risk parameter (γ^+) and the behavioural consistency parameter (β), and was implemented to enable comparisons between the groups in our data set. A few alterations to the model were made in order to fit the CCT framework. Firstly, the number of cards considered optimal (ω) was specified at the subject level and updated on every trial, rather than being constant across trials like in van Ravenzwaaij et al. 's (2011) model. Regarding the probability of encountering a loss card, we adopted a similar approach as the original model creators. We used a constant probability of 0.1 across all trials and flips during the game, based on the finding that probabilities outside the range of 0.1-0.2 lead to biassed results. As such, we assume participants use the available information at the beginning of each trial to create a representation of the optimal number of card flips, and that this representation is not updated as the probability of encountering a bad card increases. Therefore, it is the initial assumption of the optimal amount of card flips which guides participants' choices in the game. This violates the important aspect of information use in the CCT. Also, using a constant probability of encountering a loss card violates the dynamic probability structure of the CCT, in which this probability begins at 0.03 or 0.09 depending on the condition of the current trial. However, since a relatively high proportion of flips in the trials will have a probability within the recommended range, this choice is deemed feasible.

[redacted]

2.2.1 Testing group differences

(ESH) The comparisons between the two groups in our data set was done via the implementation of hierarchical Bayesian inference (HBI) two-sample t-tests in the model structure. The HBI t-test is a method for determining uncertainty of a posterior over some group parameters (Piray et al., 2019). In our case, this entailed computation of posterior estimates for the difference (δ) between group parameter means. Assessment of the t-tests was done through computation of bayesian 95% credible intervals. Additionally, Bayes Factors for the differences in group parameter means (δ_{γ} and δ_{β}) were calculated as the Savage-Dickey density ratio (Lee & Wagenmakers, 2013), and interpreted according to standards of Kass & Raftery (1995). The hierarchical model was implemented using JAGS (Plummer, 2003) with the *R2Jags* package (Masanaou & Su, 2021) in Rstudio (R Core Team, 2022). The subject level free parameters for each group were positively constrained, and gamma distributed in terms of the shape and rate of the free group parameters. The shape and the rate were later reparameterized to mean and standard deviation, for implementation of the t-tests. The gamma distribution was chosen over the Gaussian distribution used by Lee and Wagenmakers (2014).

(NMA) Deviance priors for both free parameters on a group level were specified in terms of precision. Here, Jeffreys prior, a non-informative prior distribution, was used. The group means for the two free parameters ($\mu\gamma_A$, $\mu\beta_A$, $\mu\gamma_B$, $\mu\beta_B$) were given by the opposing distances between the overall common mean for each parameter ($\mu\gamma$, $\mu\beta$) and half of the estimated difference in group means ($\delta\gamma$, $\delta\beta$) (*see equations below).* This specification is feasible because the priors for the overall common mean (μ) and differences between groups (δ) for both free parameters share the same type of distributions, namely the standard normal distributions.

The specifications for the means of risk propensity (γ) for the groups compared in the Bayesian hierarchical t-tests:

$$\mu\gamma_A \leftarrow \mu\gamma - \frac{\delta\gamma}{2}$$

$$\mu\gamma_B \leftarrow \mu\gamma + \frac{\delta\gamma}{2}$$

.

[redacted]

The means of behavioural consistency (β) for the groups compared in the Bayesian hierarchical t-tests:

$$\mu\beta_A \leftarrow \mu\beta - \frac{\delta\beta}{2}$$

$$\mu\beta_B \leftarrow \mu\beta + \frac{\delta\beta}{2}$$

The group means were log transformed, since implementation of the hierarchical t-test required these to be in linear space. The posterior sampling for all the group comparisons included four Markov chains, each with 5000 iterations and a burn-in of 1000 samples.

2.3 Hierarchical parameter estimation of CCT data with the statistical BART model

2.3.1 Data preprocessing

(LT) In order to fit the statistical BART-model to the CCT data, several steps were taken to ensure that the data was in the expected format. The data provided for the sampling must contain 4 different lists with information regarding (1) no. of pumps (y) for each participant in every given trial. (2) Information about the bursting point for that given trial, (3) the individual participant ID, and (4) the group that the participant belongs to (z). Unfortunately, the original dataset only contains information regarding burst points, in an individual trial, if a burst occurred in that trial. Since the model assumes we know the bursting point of every trial, even those that did not result in a burst, this poses a slight problem. To solve this, we set a random value for the bursting point for banked trials, i.e a value that specifies where a card flip would have resulted in a burst if the participant did not choose to bank. This value must be greater than y and equal to or smaller than 32 (max amount of cards). The values for 1,3 and 4 were extracted from the original dataset.

2.3.2 Testing group differences

(ESH) Coon and Lee (2021) present a way to extend their model for group comparisons, which uses the hierarchical approach described by Lee and Wagenmakers (2013, Chapter 8). We test for group differences, between crack users and healthy adults, in accordance with this adaptation of

the model. For group-level estimations, this approach is similar to that of the cognitive BART model described above, however, differs slightly in some aspects. First, this model doesn't feature a parameter to capture the number of cards considered optimal prior to flipping (ω), which removes the need to specify and assume an underlying burst probability structure when sampling. The model only uses the observed no. of pumps (y) and points of burst, to estimate rho and beta for each individual participant. This is extended to the group level by assuming that crack users and healthy adults belong to two different groups, both in terms of their risk propensity and behavioural consistency. These groups are formalised as truncated gaussian distributions with means $\mu\rho_1$, $\mu\beta_1$ if $z_i = 1$, and means $\mu\rho_2$, $\mu\beta_2$ if $z_i = 2$. Similarly, the standard deviations are denoted by $\sigma\rho_1$, $\sigma\beta_1$ if $z_i = 1$ and $\sigma\rho_2$, $\sigma\beta_2$ if $z_i = 2$. If the ith participant is a healthy adult $z_i = 1$, and $z_i = 2$ if the ith participant is a crack user. As with the previous model, the group means are given by the differences between them, which involves denoting a grand mean for that parameter distribution along with a difference parameter (δ):

Mean risk propensity (ρ) for group 1 and 2:

$$\mu\rho_1 \leftarrow \mu\rho - \frac{\delta\rho}{2}$$

$$\mu\rho_2 \leftarrow \mu\rho + \frac{\delta\rho}{2}$$

Mean behavioural consistency (β) for group 1 and 2:

$$\mu\beta_1 \leftarrow \mu\beta - \frac{\delta\beta}{2}$$

$$\mu\beta_2 \leftarrow \mu\beta + \frac{\delta\beta}{2}$$

(NMA) Accordingly, the posterior distribution of delta (δ) for both ρ and β will inform us on the difference between their means. By assessing the mean of the posterior distributions of δ for both parameters, and employing the Savage-Dickey method (Lee & Wagenmakers, 2013) to calculate the Bayes Factor, it is possible to examine how strong the evidence is for the group means of

both risk propensity and behavioural consistency to be different between crack users and healthy adults.

The model implementation in the original paper (Coon & Lee, 2021) assumes that a participant's behavioural consistency (β), is independent of their group. As shown above, we modified the model to contain the group-dependent distributions for the beta as well, allowing us to test for group differences in both parameters. Accordingly, the ρ and β of each participant is sampled from their respective group:

$$\rho_1 \sim Gaussian_+\left(\mu\rho_1, \frac{1}{\sigma\rho_1^2}\right) if\ z_i = 1,\ \rho_2 \sim Gaussian_+\left(\mu\rho_2, \frac{1}{\sigma\rho_2^2}\right) if\ z_i = 2$$

$$\beta_1 \sim Gaussian_+\left(\mu\beta_1, \frac{1}{\sigma\beta_1^2}\right) if\ z_i = 1,\ \beta_2 \sim Gaussian_+\left(\mu\beta_2, \frac{1}{\sigma\beta_2^2}\right) if\ z_i = 2$$

This hierarchical model was also implemented using JAGS (Plummer, 2003) with the *R2Jags* package (Masanaou & Su, 2021) in Rstudio (R Core Team, 2022). The posterior sampling for the group comparisons included four Markov chains, each with 10.000 iterations and a burn-in of 1000 samples.

2.3.3 Priors

(LT) Based on the Kluwe-Schiavon et al., (2016) experiment we set the prior for the grand means of the parameters to $\mu\rho \sim Gaussian_+(7, 1/5^2)$, $\mu\beta \sim Gaussian_+(3, 1/2^2)$. We set the variance as such, to allow for fairly large individual differences in both risk propensity and behavioural consistency. We set priors for standard deviations and the deltas like this:

- $\sigma\rho_1, \sigma\rho_2, \sim Gaussian_+(0, 1/10^2)$, $\sigma\beta_1, \sigma\beta_2, \sim Gaussian_+(0, 1/10^2)$
- $\delta\rho, \delta\beta \sim Gaussian(0, 1/10^2)$

The prior distributions for delta has means of 0, as we assume zero difference in the groups, and allow for an effect in both directions.

2.4 Assessing the efficacy of the models

(ESH) In addition to the individual models' convergence, we assessed the efficacy of both models on the basis of their ability to successfully recover known parameters, and through posterior predictives. The following sections outline our approach when doing so.

2.4.1 Parameter recovery

(ESH) In this section, parameter recovery of both the statistical BART model and cognitive BART model is examined. To do so we set up a function in Rstudio that simulates participant data based on know values of ρ and β, (γ and β for the cognitive BART model). This entails first setting up a function that simulates burst points, i.e which card would result in a burst. Since we are simulating the CCT, the function to simulate burst point simply generates a list of burst points, that is the number of trials long with randomly sampled values between 1-32. We then calculate no. of pumps (y) for a given participant, the same way the respective model does, across an arbitrary number of trials. We choose 24 trials, as this is the number of trials in the original experiment. When estimating the number of flips for a participant in the cognitive BART model, we set a probability of burst to 0.1, for the same reasons outlined earlier. The values for ρ_{true} and β_{true} were sampled from truncated gaussian distributions given by:

$$\rho_{true} \sim Gaussian_{+}(M = 7, SD = 6), \text{and } \beta_{true} \sim Gaussian_{+}(M = 3, SD = 3).$$

For both models, we simulated data for 100 participants, which entails running 100 different models, across 4 chains, with 10.000 iterations and burn in of 1000.

2.4.2 Posterior predictive checks

(NMA) We employed posterior predictives to assess the model fit of both models. This entails using the models to generate data from the participant's estimated parameters for risk propensity

[redacted]

and behavioural consistency. Simply put, if the data generated from the estimated parameters of a participant closely resembles the actual data from that participant, we can suggest that the model fit is good. Accordingly, we examine the predictive accuracy of our two models by comparing the predicted no. of card flips to the actual no. of flips across 24 trials for all 47 participants. The distribution of predicted number of card flips for each participant were calculated using their respective mean estimate for risk propensity and behavioural consistency. This approach to assess model fit was inspired by D'Alessandro et al. (2020) and Ravenzwaaij et al. (2011), who employed a similar method.

2.5 Github Repository

(ESH, NMA, LT) All code for data processing, sampling and visualisation can be found in this Github repository: [redacted]

3.0 Results

3.1 Cognitive BART model fit to hot CCT data

3.1.1 Model convergence

(LT) Model convergence was assessed through trace plot inspection, as well as assessment of the potential scale reduction factors (R-hat) for each estimate tracked in the t-test (see the appendix for trace plots and exact R-hat values for all estimates). All posterior estimates had R-hat values below or equal to 1.01, which is considered good convergence according to Vehtari et al. (2021). All posterior estimates had good convergence based on this threshold. This was also reflected in good mixing of chains for all estimates.

3.1.2 Group comparison and estimates

(ESH) Crack users and controls were found to differ in terms of behavioural consistency ($M\delta_\beta$ = *0.53,* $SD\delta_\beta$ *= 0.26,* $CI\delta_\beta$ *= 0.08-1.09*) *(visualised in Fig. 4).* With a Bayes Factor of 3.84, there is thus substantial evidence that there is indeed a difference, as interpreted according to the table provided by Kass and Raftery (1995). Through inspection of the model output (*see appendix*) we

find that the estimated mean behavioural consistency of crack users (M_{β}= *0.28,* SD_{β}=*0.04*) was slightly higher than that of controls (M_{β}= *0.17,* SD_{β}=*0.03*). No indices of a difference between the two groups were found for risk propensity (*Mδγ = -0.26 , SDδγ =0.20, CIδγ =-0.71-0.12)*. A Bayes Factor of 0.49 for this difference is an effect "*not worth more than a bare mention"* (Kass & Raftery, 1995). Further inspection of the model output could suggest that controls (*Mγ =1.96 , SDγ =0.24*) are slightly more risk seeking than crack users (*Mγ =1.53 , SDγ =0.23*). The output for all model estimates can be found in the appendix.

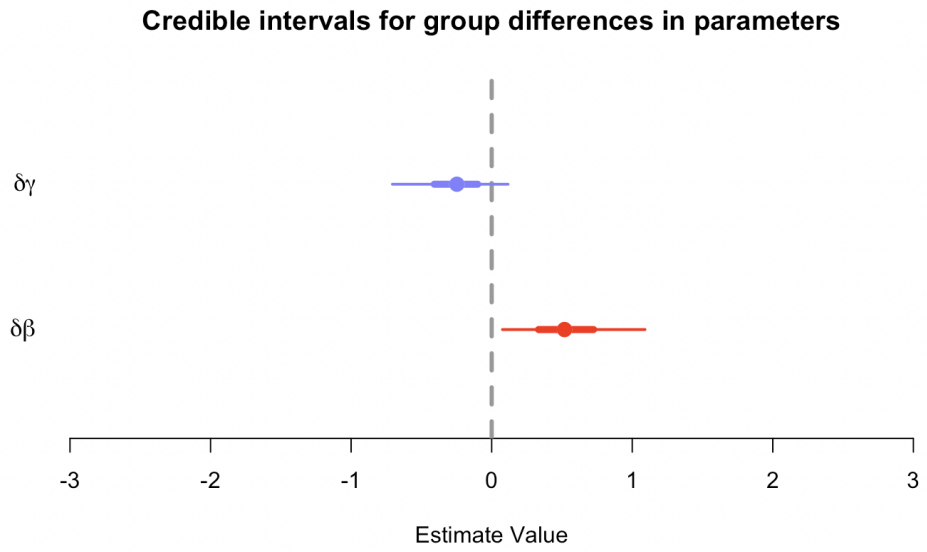


Figure 4: Caterpillar plot showing the output of the Bayesian t-test. The plot shows the 95% and 60% Bayesian credible intervals for the difference (δ) in posterior mean risk propensity (γ), and in posterior mean behavioural consistency (β) between crack users and controls. Wide, thin lines indicate 95% credible intervals, and narrow, thick lines indicate 60% credible intervals. Points indicate posterior medians. Parameters where both credible intervals overlap with 0 are illustrated with open circles in combination with light colour. Lastly, a darker shade combined with a filled circle is shown for parameters where neither of the two credible intervals overlap with 0. The figure thus indicates that, with 95% probability, the difference in behavioural consistency between controls and crack users is non-zero. Likewise, with 95% probability, the posterior for the estimated difference in risk propensity between the groups does include zero.

3.2 Statistical BART to CCT

3.2.1 Model Convergence

(NMA) Model convergence was again assessed through trace plot inspection, as well as assessment of the potential scale reduction factors (R-hat) for each estimate. From visual inspection of the traceplots for delta of beta and mean of beta for group 1, we would consider the convergence for these parameters sub-optimal. The summary output reveals that all posterior estimates had R-hat values below or equal to 1.01, expect these two, who report an rhat of 1.012. We take this into account when interpreting the results.

3.2.2 Group comparison and estimates

(NMA) The results of the group level differences are shown in *Figure 5*. The mean of $\delta\rho$ is 3.79 with a 95% credible interval of (-10.13, 17.12), indicating that there is no difference between crack users and controls in terms of risk propensity. The mean of $\delta\beta$ is 2.27 with a 95% credible interval of (-4.99, 8.13), also indicating a lack of any substantial difference between the two groups. For $\delta\rho$ the Bayes Factor is 1.19, and for $\delta\beta$ the Bayes Factor is 2.04, which both are effects “not worth a bare mention” according to the table provided by Kass & Raftery (1995). Despite the lack of evidence for a difference between the groups, the posterior distributions in *Fig. 5* are indicative of both a higher risk propensity (ρ) and behavioural consistency for crack users.

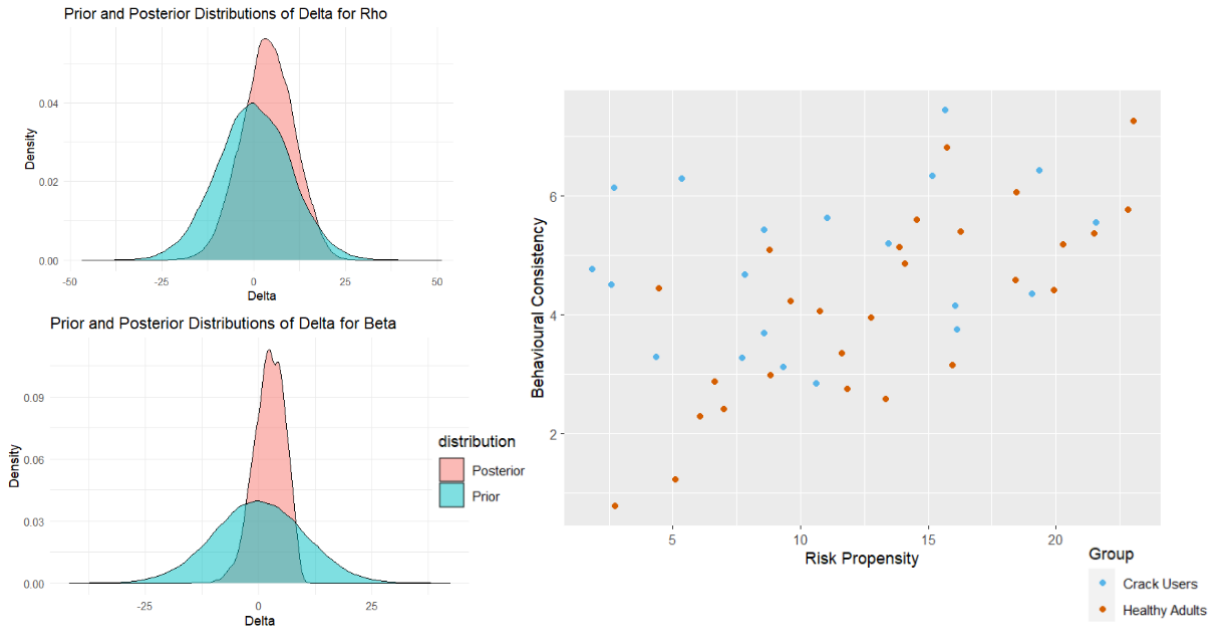


Figure 5: Scatter plot of posterior means of estimated rho and beta measures for each participant, along with prior and posterior distributions of delta for each measure. A displacement of the posterior distribution towards the right is indicative of higher posterior means for both free parameters for the crack group (denoted as group 2 in the analysis.)

3.3 Results of parameter recovery

(LT) Results of the parameter recovery for the two models can be seen in *Fig. 6*. For the statistical BART, the RMSE for the ρ recovery was 1.41 and 1.07 for β. For the cognitive BART model the RMSE was deemed irrelevant to calculate, as the parameter recovery for the single participant model was not recovering the parameters correctly. The parameters recovered from this model seems vastly different than the ones estimated by the group-level model for the same ranges of rho and beta, a finding that will be elaborated upon in the discussion section.

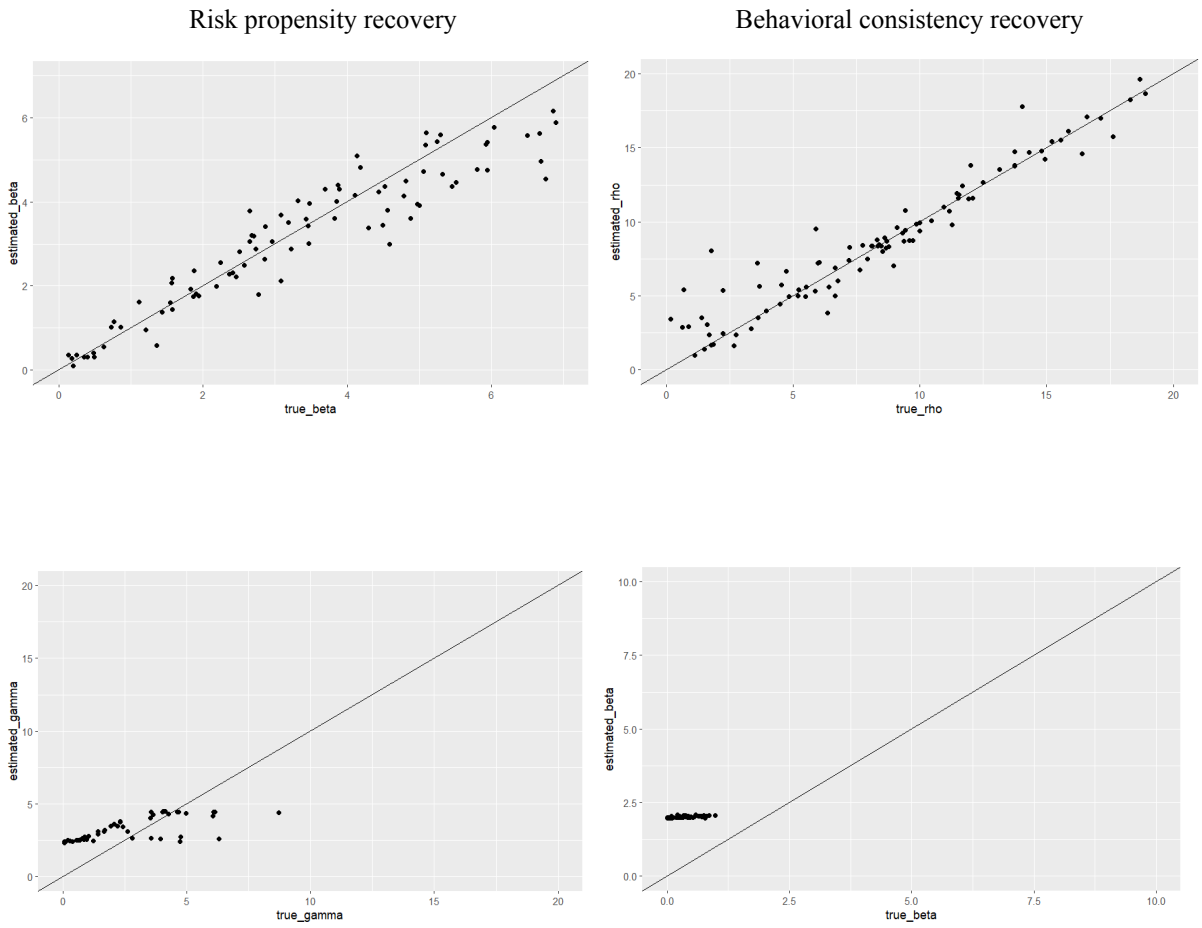
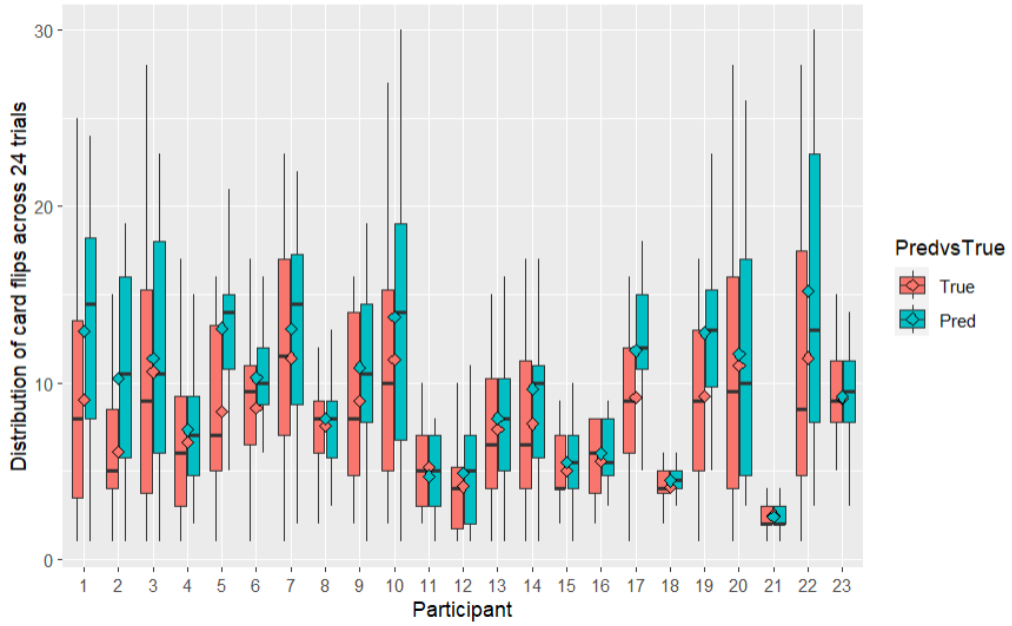


Figure 6. The relationship between the true parameters and the mean of the estimated parameter distributions, for the CCT-like simulation of 100 participants. First row of plots show recovery for the statistical BART model, and the second row for the cognitive BART .

3.4 Results of Posterior predictive checks

(ESH) The predicted distribution of no. of flips for 24 trials along with the true distribution for each participant is visualised in *Fig. 7 a+b* for the statistical BART model.

(a)



(b)

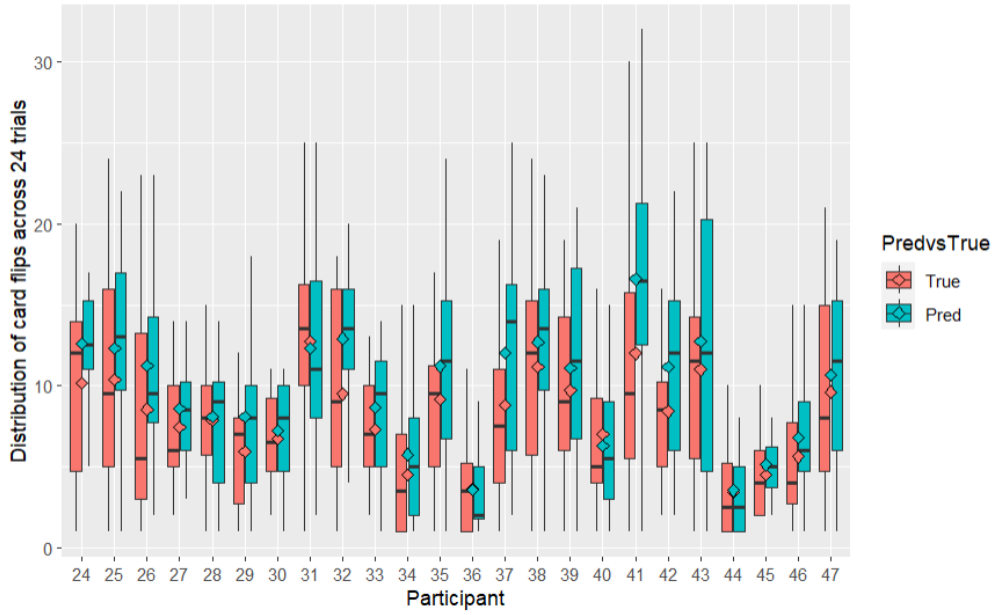
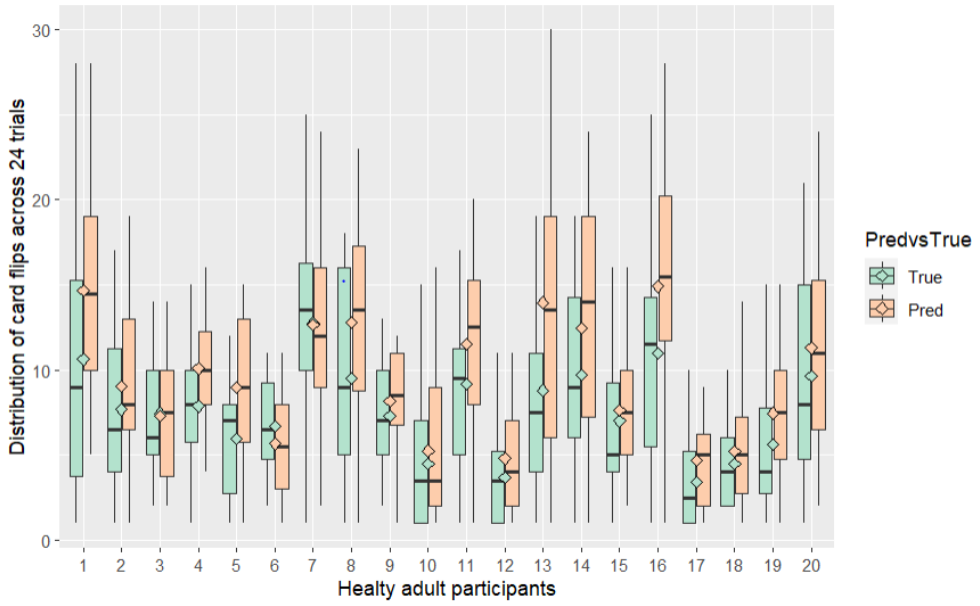


Figure 7: Grouped boxplots showing the predicted and true distribution of card flips across 24 trials for participants 1-23 (a) and 24-47 (b), for the statistical BART model. The diamond shapes represent mean no. of flips.

(NMA) The predicted distribution of no. of flips for 24 trials along with the true distribution for each participant is visualised in *Fig. 8* a+b, for the cognitive BART model.

(a)



(b)

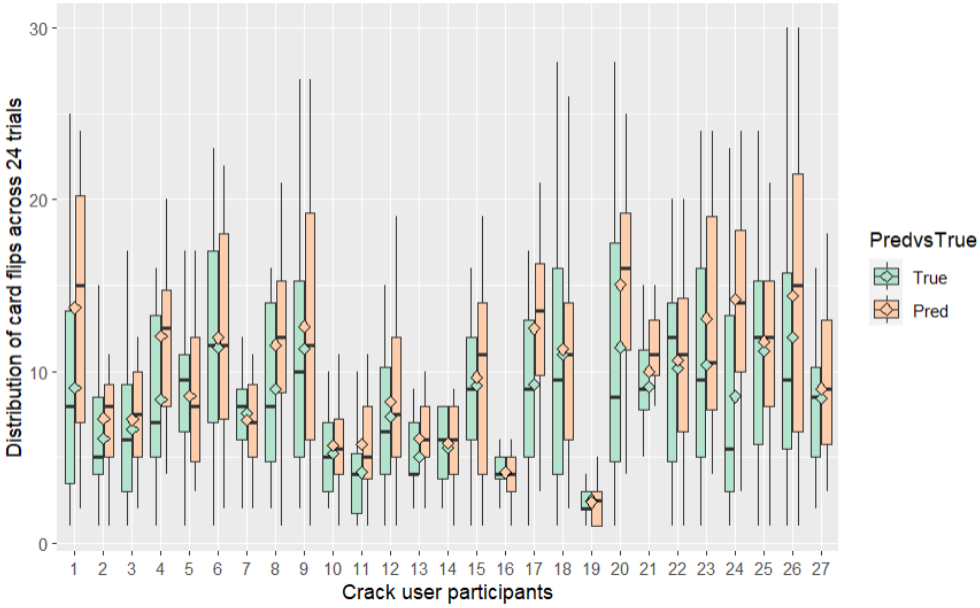


Figure 8: Grouped boxplots showing the predicted and true distribution of card flips across 24 trials for healthy adult participants (a) and crack users (b), for the cognitive BART model. The diamond shapes represent mean no. of flips.

4.0 Discussion

4.1 Group differences

(LT) In this paper, we applied and compared the performance of two different BART-models on data from the hot version of the CCT. The purpose of this approach was twofold: Firstly, we wanted to investigate the differences in risk propensity in two different population groups, namely crack users and healthy adults. Secondly, we wished to investigate the applicability and generalisability of different methods for analysing data from different risk measures applied in Cognitive Science. We start the discussion of this paper by presenting and nuancing our findings, compare the performance of the two model approaches and address possible explanations behind these. In the cognitive BART model, no differences in risk propensity between crack users and controls was found. This finding is in concordance with the paper from which we obtained our data, where it was concluded that crack cocaine users modulate their risk behaviour in response to environmental feedback regarding their performance in the CCT (Kluwe-Schiavon et al., 2016). In terms of behavioural consistency, it is slightly surprising that controls were estimated to have lower behavioural consistency than crack users. In the original paper, controls portrayed stable choice patterns that were based on use of information directly available in the task. (ESH) Cocaine addicts on the other hand were found to guide their choices according to reward and punishment contingencies associated with their behaviour. We would thus have expected that crack cocaine users' estimated behavioural consistency would be lower than that of controls. When it comes to the statistical BART model, no differences between the groups were found. However, the trends in the estimated posterior distributions for both risk propensity and behavioural consistency hint towards slightly higher estimates in the dependent cohort, but with the evidence being too sparse to conclude anything decisive. What we can discuss, however, is the fact that the two model approaches both identify no difference in risk propensity between the two groups in question. Such a finding is in concordance with evidence of variable effects of cocaine use on performance on the BART or another popular risk task called the Iowa Gambling task (as referred to in Wittwer et al., 2016). Possible reasons behind different estimates for behavioural consistency in the two approaches applied in this paper can further be understood in the light of the model's performance in contrast to each other.

4.2 Model efficacy

4.2.1 Parameter recovery

(NMA) When assessing the parameter recovery of the cognitive BART model, it becomes clear that the single participant model we used to recover the parameters wasn't implemented correctly, it would have been more beneficial to study parameter recovery for the hierarchical model, as the posterior predictive checks show that this model "recovers" parameters like expected. However, generating participant data in a group-like structure for parameter recovery was deemed out of the scope of this particular paper. The relationship between the true and estimated parameters for the statistical BART model, visualised in fig. 6, show that the parameter estimates of the model are generally close to the sampled true values. This suggests that our model successfully recovers parameters in the expected manner. It must be said that such parameter recovery results solely provide evidence that this model is correctly implemented, and that 24 trials in a Hot-CCT like simulation is enough to recover these parameters fairly well. This is a promising finding, as it proves the adaptation of the experiment data does not alter the models ability to estimate risk propensity and behavioural consistency of a participant. However, the successful recovery does not show whether the estimated parameters are useful for modelling the behaviour of participants. This can be examined further from the results of the posterior predictive checks.

4.2.2 Posterior predictive checks

(LT) We assess the quality of the model fits by examining the overlap of the predicted and true distribution of pumps for each participant. A way of investigating how well the risk propensity estimate of a participant predicts no. of pumps, could be to look at the difference in mean no. of pumps for predicted vs true data. As is apparent for both models, these predictions are sometimes very accurate, but fall short in other cases. Both models often overestimate how many flips a participant is likely to make. For behavioural consistency predictions we can assess the whiskers and the interquartile range of the distributions. Like the means, these align fairly well, however the inconsistency of a participant is also often overestimated slightly. These are valuable insights, as they highlight the need for further refinement and optimization of the models if they are to be used for CCT data. Based on these posterior predictive checks we cannot suggest that the model fits are good, as a better alignment of distributions would be preferred. That being said, we still deem the results of the posterior predictive checks promising for the utility of these BART

models for hot-CCT. This is because both models seem to reliably estimate parameters that to some extent can predict participant behaviour.

(ESH) It is important to note that there are some limitations to our approach for posterior predictive checks, as these predicted distributions are solely generated from the mean estimate for the 2 parameters. To get a more sound overview of the predictive power of the models it would have been beneficial to sample e.g 1000 values of risk propensity and behavioural consistency from the posterior distributions for each participant, and generate pump distributions for each parameter pair. The density of the means of these distributions could then be compared to the confidence intervals and means of the actual number of pumps for each participant. This is how predictive power is assessed in the Ravenswaaij et al. (2011) paper, and would be good to employ in a future study of model generalisation to ensure a more statistically valid assessment of our model fits.

4.2.3 Model comparison

(NMA) Based solely on the results of the parameter recovery and the posterior predictive checks, it is difficult to assess which model best suits hot-CCT data. It is interesting that the predictive strength of the cognitive BART model is similar to that of statistical BART when the latter assumes that participants base their decision-making on a burst probability of 0.1. For future research it would be interesting to test whether other burst probabilities in that range have better predictive power, and also if the model could be implemented with a dynamic burst probability structure. Utilising a dynamic burst probability structure, in the cognitive BART model, would entail updating omega for every choice based on a probability given by how many cards are left to flip. Since the statistical BART model does not assume that participants account for burst probability in their decision-making, we would argue that this is theoretically the best model to use for hot-CCT parameter estimation. Based on the posterior predictives, we assume that if a larger group level effect had been present in the analysed hot-CCT data, both models would have been likely to capture it in the posterior distributions of the deltas.

[redacted]

4.3.1 Methodological limitations

(LT) Additional limitations in this study must be addressed, and should be taken into consideration when interpreting or building upon the results presented. These include issues specifically related to the model specifications, to measuring risk propensity via an experimental paradigm, as well as the question of how one can generalise findings from an experimental setting to a real life environment in an informative manner.

First and foremost, several of the choices made when fitting the two BART models to the hot CCT data reduce the theoretical interpretability of our results. Starting out with the cognitive BART model, there is a major drawback that the model does not account for how information is used in the game. As presented in the theory section, decisions made under risk and uncertainty are most likely highly dependent on factors such as different combinations of loss and gain values in the experiment. By not including these aspects in the model, we assume that the cognitive processes involved in the BART and the CCT are exactly the same, which findings have suggested they are not (Buelow & Blaine, 2015). In case that the two tasks really do touch upon different aspects of the decision making process, applying a cognitive model designed for the BART will produce inaccurate predictions regarding CCT behaviour. Secondly, assuming a constant probability of encountering a loss card in the hot CCT is directly wrong, and thus a poor reflection of the actual task environment. Although this choice made sense in the light of the purposes of the current paper, it cannot be ignored that such a choice further reduces the interpretability and applicability of the current model. Lastly, our use of completely uninformative priors when fitting the cognitive BART model is a potential problem. By using informed priors, there is a chance that we could have obtained better justified and interpretable results. Also, considering that informed priors were used in the statistical BART model, using the same approach in the cognitive BART model might have been beneficial for increasing comparability between the models.

4.3.2 Theoretical limitations

(NMA) Apart from the methodological limitations, it is relevant to address which theoretical limitations the current paper faces. One important constraint is that transition of a measure obtained in the laboratory is tricky to generalise to real-life (Holleman et al., 2020). This is grounded in the fact that creating an experimental environment in which participants really feel

that something important is at stake is difficult, considering the many ethical- and economic constraints that govern such research (Ert & Yechiam, 2010). Further, as previously touched upon, it is a hard task to discern risk propensity as a latent measure from similar, unobservable concepts such as loss aversion and risk aversion. This issue is inherent in all measurement methods, and is typically aimed to be addressed in the creation of different experimental designs. Distinctions made in the experimental design must therefore be sufficiently accounted for when attempting to model and analyse data.Since the objective of our approach was to evaluate the applicability of models designed for estimating risk propensity in BART for similar tasks, such design-related distinctions were not taken into account. In contrast, we make several assumptions to be able to model data of different structures, which might even further reduce the transparency of the concept modelled. This is especially relevant to consider in our case, as Pedroni et al. (2017) found substantial differences between how risk propensity is represented in BART and hot CCT. Also, there is a lack of consensus as to how similar BART and hot CCT actually are in terms of decision making processes and risk propensity, with some researchers suggesting that they measure distinct decision making processes (Buelow & Blaine, 2015). When evaluating the obtained measures which we call risk propensity, it is therefore important that we keep in mind that these measures might be related to risk- and loss aversion, as well.

4.5 Future prospects

(ESH) Our analysis revealed that the modified BART models, while informative in some ways, were not sufficient to fully capture the complexities inherent in CCT data. Nonetheless, the models show promise for future research, as there is significant room for improvement and further accounting for the unique structure of CCT data. Further refinements to the models, such as the ones suggested above, may allow for a more robust and reliable estimation of the underlying decision-making parameters in CCT studies. Lastly, future research efforts on this matter should consider whether creating a general model for analysing risk data, as in the current paper, actually makes sense. In this paper, we find that existing models can be generalised to similar experimental tasks to a certain extent.

(NMA) However, the long list of limitations accompanied by such an approach leads to a questionable degree of utility of such a model. A general model for analysing experimental risk data can be useful when the goal is to compare such data across studies and experimental

paradigms. This way, one might be able to capture general tendencies in a more coherent way than when using very different modelling approaches. On the other hand, if the goal is to provide an in-depth interpretation of the decision making behaviour in a specific experimental task, a custom-made model will more likely capture the relevant cognitive processes involved in that task, given that it is difficult to create a general model without making assumptions that violates certain aspects of different experimental tasks.

5.0 Conclusion

(ESH, NMA & LT) This paper aimed to investigate the utility of two Bayesian BART models when applied to data derived from the hot-CCT. The efficacy of applying these models to the CCT was investigated by fitting each model to data from an experiment with crack users and healthy adults. With the exception of a slight group-level difference in beta for the cognitive BART model, the posterior distributions of the deltas suggested that no group-level difference of risk propensity and behavioural consistency could be inferred. The validity of this finding was supported by a parameter recovery and posterior predictive check for both models. These model efficacy tests also served to examine the overall generalizability of BART models to hot-CCT data, and suggests that, though the results are promising, further adaptation of such models are required to generalise well across sequential risk-taking tasks. Finally, it was discussed which modifications could be made, and the merits and shortcomings that they each entail. Overall, this study contributes to the ongoing effort to develop more effective and appropriate models for measuring risk-taking behaviour in the CCT and other related tasks.

References

Agarwal, R. (2022, April 12). *MCMC Intuition for Everyone*. Medium. https://towardsdatascience.com/mcmc-intuition-for-everyone-5ae79fff22b1

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov, & F. Csaki (Eds.), Proceedings of the 2nd International Symposium on Information Theory (pp. 267-281). Budapest Akademiai Kiado. - References—Scientific Research Publishing*.

Aven, T. (2012). The risk concept—Historical and recent development trends. *Reliability Engineering & System Safety*, *99*, 33–44. https://doi.org/10.1016/j.ress.2011.11.006

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*(1–3), 7–15. https://doi.org/10.1016/0010-0277(94)90018-3

Bishara, A. J., Pleskac, T. J., Fridberg, D. J., Yechiam, E., Lucas, J., Busemeyer, J. R., Finn, P. R., & Stout, J. C. (2009). Similar processes despite divergent behavior in two commonly used measures of risky decision making. *Journal of Behavioral Decision Making*, *22*(4), 435–454. https://doi.org/10.1002/bdm.641

Blais, A.-R., & Weber, E. U. (2006). *A Domain-Specific Risk-Taking (DOSPERT) Scale for Adult Populations* (SSRN Scholarly Paper No. 1301089). https://papers.ssrn.com/abstract=1301089

Buelow, M. T., & Barnhart, W. R. (2018). Test–Retest Reliability of Common Behavioral Decision Making Tasks. *Archives of Clinical Neuropsychology*, *33*(1), 125–129. https://doi.org/10.1093/arclin/acx038

Buelow, M. T., & Blaine, A. L. (2015). The assessment of risky decision making: A factor analysis of performance on the Iowa Gambling Task, Balloon Analogue Risk Task, and Columbia Card Task. *Psychological Assessment*, *27*(3), 777–785. https://doi.org/10.1037/a0038622

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Coon, J., & Lee, M. D. (2022). A Bayesian method for measuring risk propensity in the Balloon Analogue Risk Task. *Behavior Research Methods*, *54*(2), 1010–1026. https://doi.org/10.3758/s13428-021-01634-1

Dang, J., King, K. M., & Inzlicht, M. (2020). Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends in Cognitive Sciences*, *24*(4), 267–269. https://doi.org/10.1016/j.tics.2020.01.007

D'Alessandro, M., Gallitto, G., Greco, A., & Lombardi, L. (2020). A joint modelling approach to analyze risky decisions by means of diffusion tensor imaging and behavioural data. Brain sciences, 10(3), 138.

Ert, E., & Yechiam, E. (2010). Consistent constructs in individuals' risk taking in decisions from experience. *Acta Psychologica*, *134*(2), 225–232. https://doi.org/10.1016/j.actpsy.2010.02.003

Figner, B., Mackinlay, R. J., Wilkening, F., & Weber, E. U. (2009). Affective and deliberative processes in risky choice: Age differences in risk taking in the Columbia Card Task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 709–730. http://dx.doi.org/10.1037/a0014983

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), e1701381. https://doi.org/10.1126/sciadv.1701381

Gelfand, A. E. (2000). Gibbs Sampling. *Journal of the American Statistical Association*, *95*(452), 1300–1304. https://doi.org/10.1080/01621459.2000.10474335

Hanoch, Y., Johnson, J. G., & Wilke, A. (2006). Domain Specificity in Experimental Measures and Participant Recruitment: An Application to Risk-Taking Behavior. *Psychological Science*, *17*(4), 300–304. https://doi.org/10.1111/j.1467-9280.2006.01702.x

Highhouse, S., Wang, Y., & Zhang, D. C. (2022). Is risk propensity unique from the big five factors of personality? A meta-analytic investigation. *Journal of Research in Personality*, *98*, 104206. https://doi.org/10.1016/j.jrp.2022.104206

Holleman, G. A., Hooge, I. T. C., Kemner, C., & Hessels, R. S. (2020). The 'Real-World Approach' and Its Problems: A Critique of the Term Ecological Validity. *Frontiers in Psychology*, *11*. https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00721

Huang, Y., Wood, S., Berger, D., & Hanoch, Y. (2013). Risky choice in younger versus older adults: Affective context matters. *Judgment and Decision Making*, *8*(2), 9.

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and Change in Risk-Taking Propensity Across the Adult Lifespan. *Journal of Personality and Social Psychology*, *111*(3), 430–450. https://doi.org/10.1037/pspp0000090

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, *246*, 160–173.

https://doi.org/10.1038/scientificamerican0182-160
Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572
Kennedy, L., Simpson, D., & Gelman, A. (2019). *The experiment is just as important as the likelihood in understanding the prior: A cautionary note on robust cognitive modelling* (arXiv:1905.10341). arXiv. http://arxiv.org/abs/1905.10341
Klaus, F., Chumbley, J. R., Seifritz, E., Kaiser, S., & Hartmann-Riemer, M. (2020). Loss Aversion and Risk Aversion in Non-Clinical Negative Symptoms and Hypomania. *Frontiers in Psychiatry*, *11*, 574131. https://doi.org/10.3389/fpsyt.2020.574131
Kluwe-Schiavon, B., Viola, T. W., Sanvicente-Vieira, B., Pezzi, J. C., & Grassi-Oliveira, R. (2016a). Similarities between adult female crack cocaine users and adolescents in risky decision-making scenarios. *Journal of Clinical and Experimental Neuropsychology*, *38*(7), 795–810. https://doi.org/10.1080/13803395.2016.1167171
Kluwe-Schiavon, B., Viola, T. W., Sanvicente-Vieira, B., Pezzi, J. C., & Grassi-Oliveira, R. (2016b). Similarities between adult female crack cocaine users and adolescents in risky decision-making scenarios. *Journal of Clinical and Experimental Neuropsychology*, *38*(7), 795–810. https://doi.org/10.1080/13803395.2016.1167171
Koffarnus, M. N., & Kaplan, B. A. (2018). Clinical models of decision making in addiction. *Pharmacology, Biochemistry, and Behavior*, *164*, 71–83. https://doi.org/10.1016/j.pbb.2017.08.010
Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C. W. (2014). Individual Differences in Risky Decision Making: A Meta-analysis of Sensation Seeking and Impulsivity with the Balloon Analogue Risk Task. *Journal of Behavioral Decision Making*, *27*(1), 20–36. https://doi.org/10.1002/bdm.1784
Lee, M. D., & Wagenmakers, E.-J. (2013). Bayesian Cognitive Modeling: A Practical Course. *Cambridge: Cambridge University Press*, 123.
Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84. http://dx.doi.org/10.1037/1076-898X.8.2.75
Machina, M. J. (1987). Choice under Uncertainty: Problems Solved and Unsolved. *Journal of Economic Perspectives*, *1*(1), 121–154. https://doi.org/10.1257/jep.1.1.121
Masanaou, Y., & Su, Y.-S. (2021). *R2jags: Using R to Rund 'JAGS'. R package version 0.7-1.* https://CRAN.R-project.org/package=R2jags
Meertens, R. M., & Lion, R. (2008). Measuring an Individual's Tendency to Take Risks: The Risk Propensity Scale1. *Journal of Applied Social Psychology*, *38*(6), 1506–1520. https://doi.org/10.1111/j.1559-1816.2008.00357.x
Nemeth, G. (2009). *The Impact of High Risk Propensity on Lifestyle and Consumption Behaviors*.
Nicholson, N., Fenton-O'Creevy, M., Soane, E., & Willman, P. (2001). Risk Propensity and Personality. *Social Research*, *8*.
O'Doherty, J. P., Cockburn, J., & Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annual Review of Psychology*, *68*(1), 73–100. https://doi.org/10.1146/annurev-psych-010416-044216
Pedroni, A., Frey, R., Bruhin, A., Dutilh, G., Hertwig, R., & Rieskamp, J. (2017). The risk elicitation puzzle. *Nature Human Behaviour*, *1*(11), Article 11. https://doi.org/10.1038/s41562-017-0219-x
Penolazzi, B., Gremigni, P., & Russo, P. M. (2012). Impulsivity and Reward Sensitivity differentially influence affective and deliberative risky decision making. *Personality and Individual Differences*, *53*(5), 655–659. https://doi.org/10.1016/j.paid.2012.05.018
Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Computational Biology*, *15*(6), e1007043. https://doi.org/10.1371/journal.pcbi.1007043
Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks. *Experimental and Clinical Psychopharmacology*, *16*, 555–564. https://doi.org/10.1037/a0014245
Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Working Papers*, 8.
R Core Team, . (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/
Ross, K. (2022). *10.1 Introduction to JAGS | An Introduction to Bayesian Reasoning and Methods*.

https://bookdown.org/kevin_davisross/bayesian-reasoning-and-methods/introduction-to-jags.html
Roy, V. (2019). *Convergence diagnostics for Markov chain Monte Carlo* (arXiv:1909.11827). arXiv. http://arxiv.org/abs/1909.11827
van Ravenzwaaij, D., Dutilh, G., & Wagenmakers, E.-J. (2011). Cognitive model decomposition of the BART: Assessment and application. *Journal of Mathematical Psychology*, *55*(1), 94–105. https://doi.org/10.1016/j.jmp.2010.08.010
Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Analysis*, *16*(2). https://doi.org/10.1214/20-BA1221
von Helversen, B., & Rieskamp, J. (2013, August 1). *Does the influence of stress on financial risk taking depend on the riskiness of the decision?*
Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling Behavior in a Clinically Diagnostic Sequential Risk-Taking Task. *Psychological Review*, *112*(4), 862–880. http://dx.doi.org/10.1037/0033-295X.112.4.862
Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414
Wittwer, A., Hulka, L. M., Heinimann, H. R., Vonmoos, M., & Quednow, B. B. (2016). Risky Decisions in a Lottery Task Are Associated with an Increase of Cocaine Use. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00640
Wright, R. J., & Rakow, T. (2017). Don't sweat it: Re-examining the somatic marker hypothesis using variants of the Balloon Analogue Risk Task. *Decision*, *4*, 52–65. https://doi.org/10.1037/dec0000055
Yildirim, I. (n.d.). *Bayesian Inference: Gibbs Sampling*.

Appendix

Model outputs - Cognitive BART model

Controls vs. Crack Users

- The A-group corresponds to controls
- The B-group corresponds to crack users
- Delta represents the difference in posterior means between the two groups compared

```
Inference for Bugs model at "constant_p.txt", fit using jags,
 4 chains, each with 5000 iterations (first 1000 discarded)
 n.sims = 16000 iterations saved
              mu.vect sd.vect     2.5%      25%      50%      75%    97.5%  Rhat n.eff
A.mu_beta       0.168   0.034    0.099    0.147    0.169    0.191    0.232 1.001  5900
A.mu_gamma      1.964   0.237    1.474    1.817    1.971    2.118    2.416 1.002  1700
B.mu_beta       0.284   0.037    0.204    0.261    0.286    0.308    0.352 1.002  2400
B.mu_gamma      1.525   0.233    1.007    1.389    1.542    1.681    1.937 1.004  1200
delta_beta      0.534   0.256    0.077    0.368    0.518    0.679    1.091 1.002  3000
delta_gamma    -0.258   0.204   -0.706   -0.373   -0.247   -0.129    0.116 1.004   740
mu_beta_log    -1.537   0.132   -1.839   -1.612   -1.521   -1.447   -1.321 1.001  6200
mu_gamma_log    0.538   0.107    0.291    0.479    0.551    0.611    0.715 1.001  4800
deviance     3079.506  16.579 3048.937 3068.005 3078.952 3090.303 3113.562 1.003  1000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule, pD = var(deviance)/2)
pD = 137.1 and DIC = 3216.6
DIC is an estimate of expected predictive error (lower deviance is better).
```

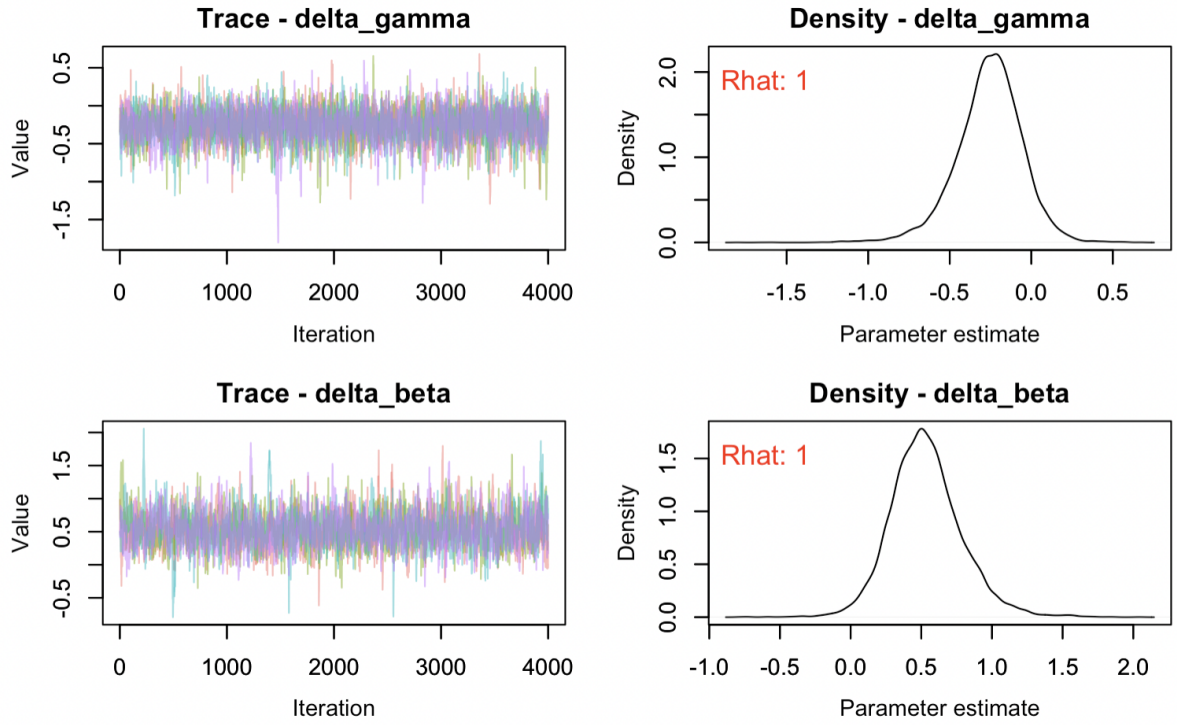
	mean <dbl>	sd <dbl>	2.5% <dbl>	50% <dbl>	97.5% <dbl>	Rhat <dbl>	n.eff <dbl>
A.mu_gamma	1.96	0.24	1.47	1.97	2.42	1.00	1555
A.mu_beta	0.17	0.03	0.10	0.17	0.23	1.01	1222
B.mu_gamma	1.53	0.23	1.01	1.54	1.94	1.00	2219
B.mu_beta	0.28	0.04	0.20	0.29	0.35	1.00	1486
delta_gamma	-0.26	0.20	-0.71	-0.25	0.12	1.00	2164
delta_beta	0.53	0.26	0.08	0.52	1.09	1.00	1345

Figure 9: The figure displays the estimated parameter values for both groups. Here, group A refers to controls, while group B refers to crack users.

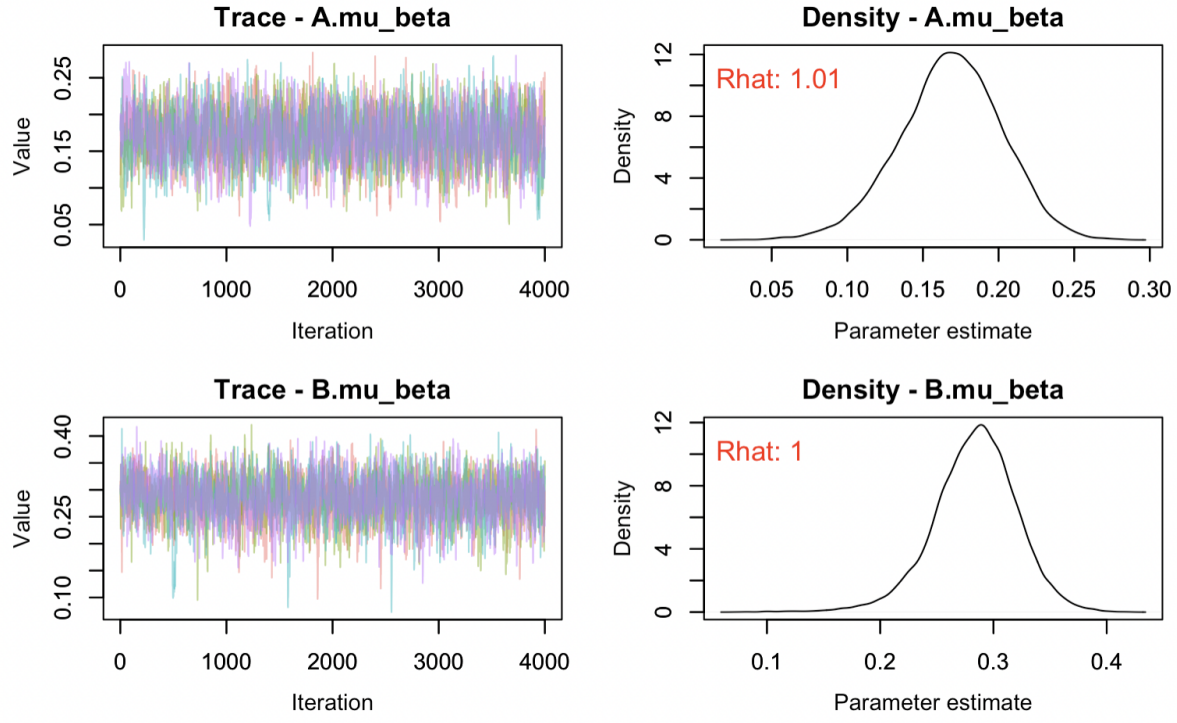
Convergence plots

Controls vs. Crack Users

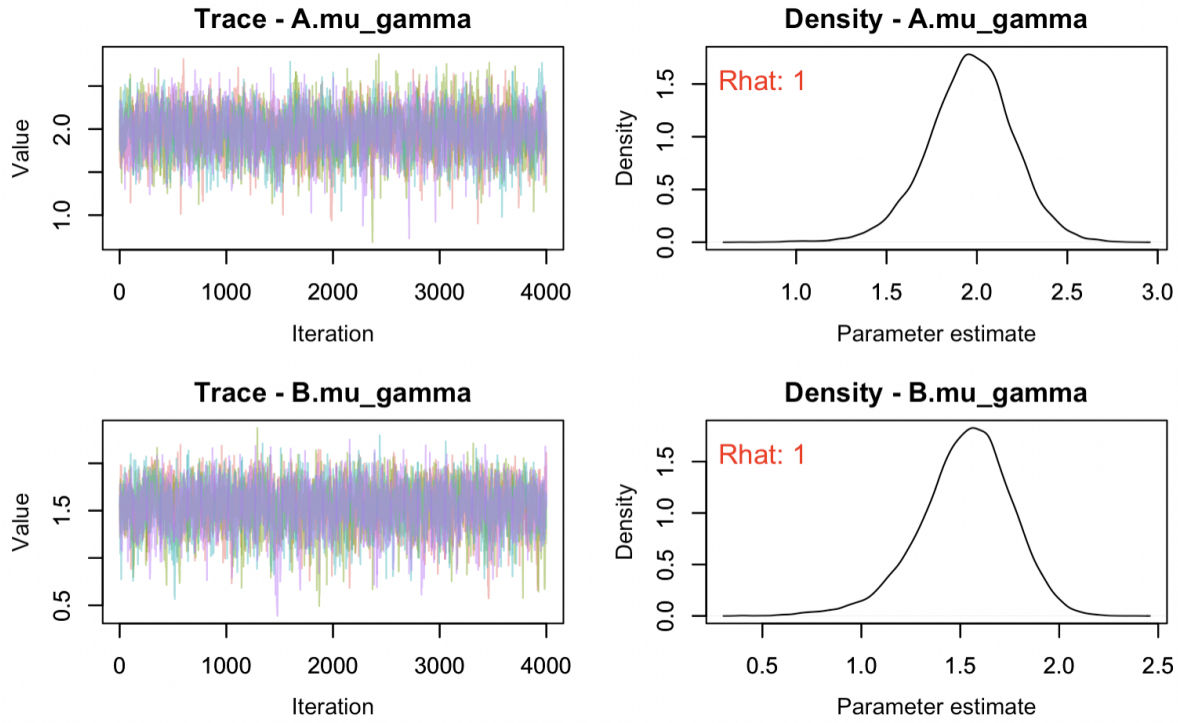
Delta



Beta (A: Controls, B: Crack Users)



Gamma (A: Controls, B: Crack Users)

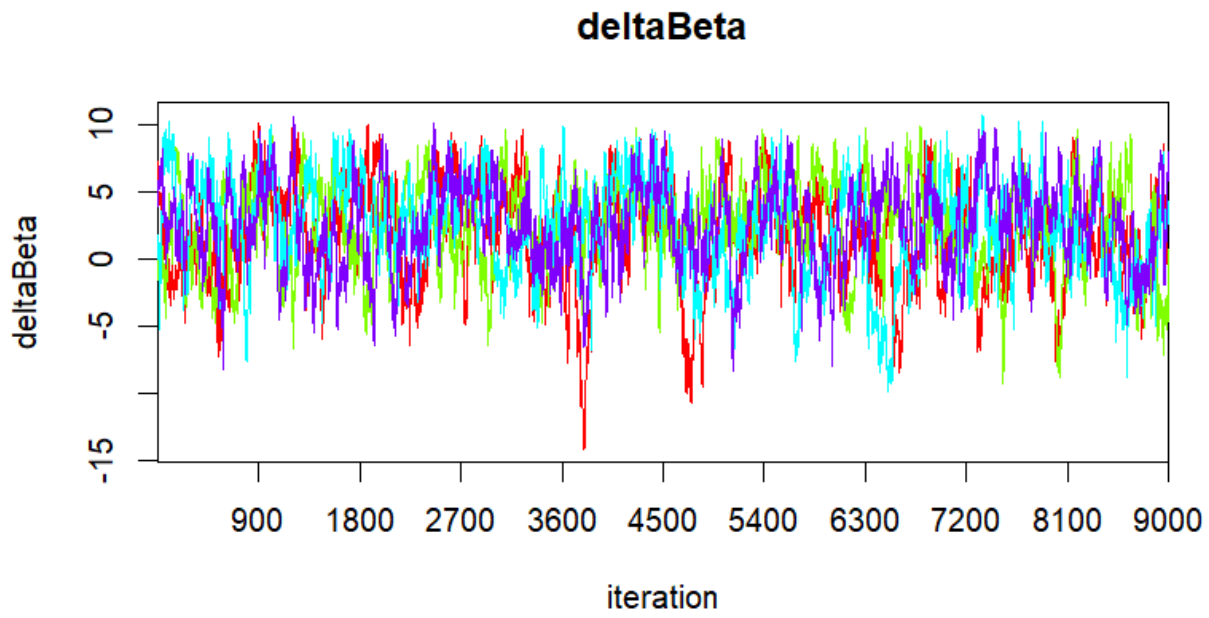


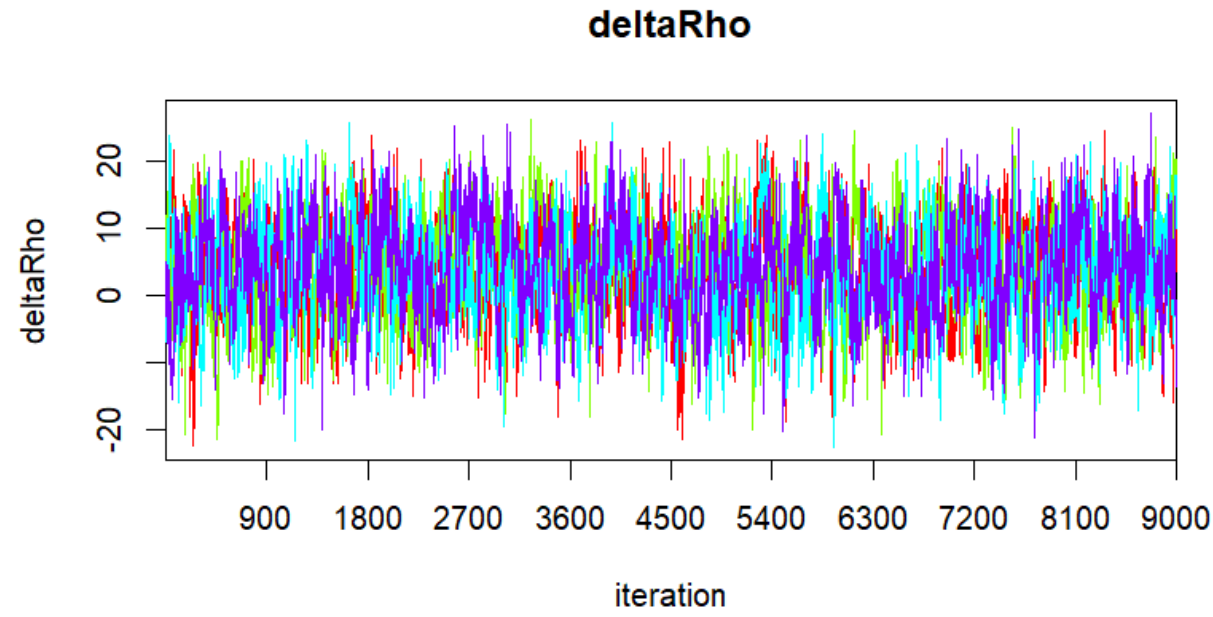
Summary of model outputs for statistical BART

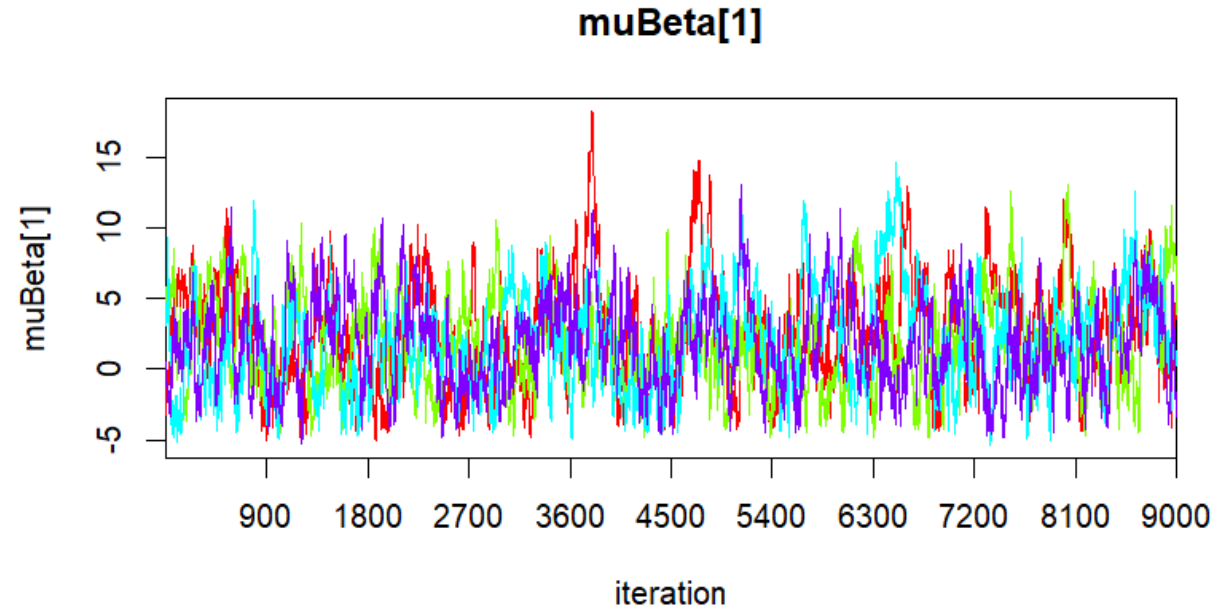
	Mean	SD	2.5%	25%	50%	75%	97,5%	Rhat	n.eff
deltaBeta	2.27278	3.4364	-4.99034	0.02843	2.47714	4.80960	8.1392	1.012	320
deltaRho	3.79080	7.0324	-10.1330	-0.91481	3.85053	8.70837	17.123	1.003	1000
muBeta[1]	2.07453	3.3781	-3.65865	-0.42782	1.86215	4.30681	9.2135	1.012	330
muBeta[2]	4.34732	0.3858	3.58976	4.11930	4.35402	4.59032	5.0786	1.001	2800

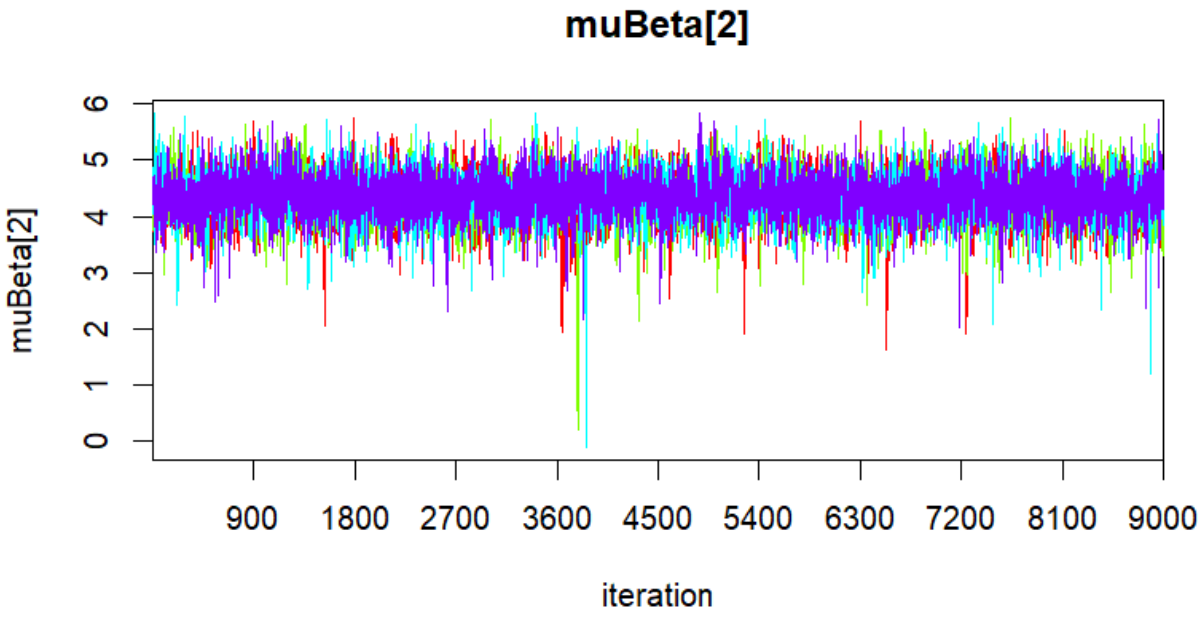
muRho[1]	7.06715	6.8023	-5.78301	2.33113	6.98107	11.6234	20.672	1.003	970
muRho[2]	10.8579	1.6899	6.78139	10.0727	11.1004	11.9543	13.400	1.001	14000

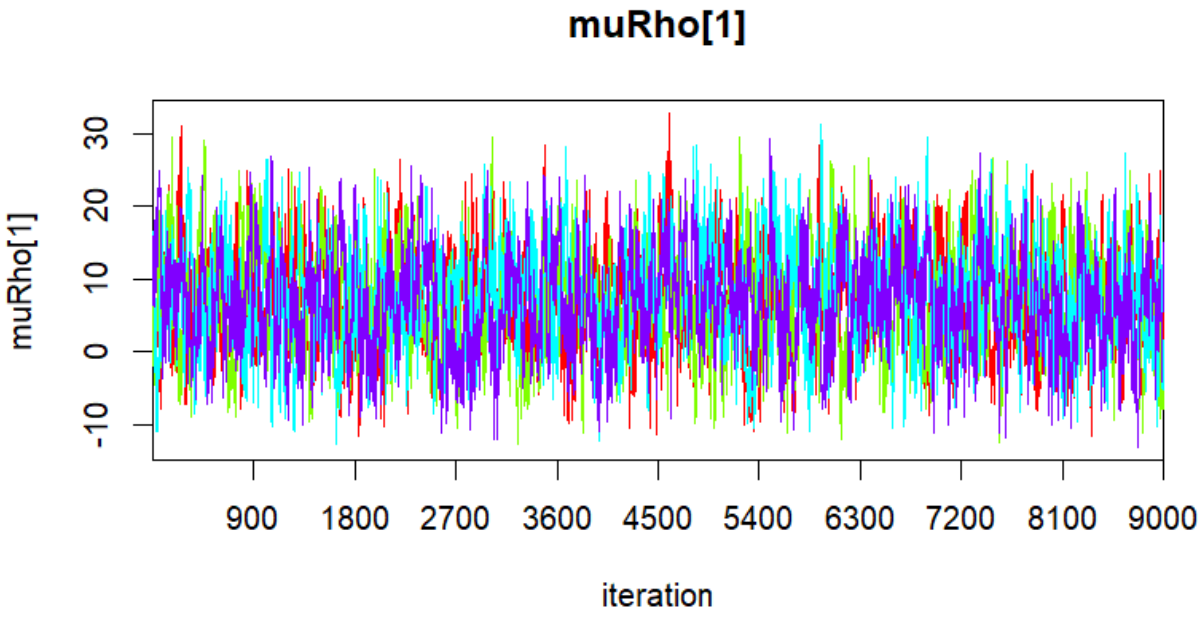
Traceplots for statistical BART



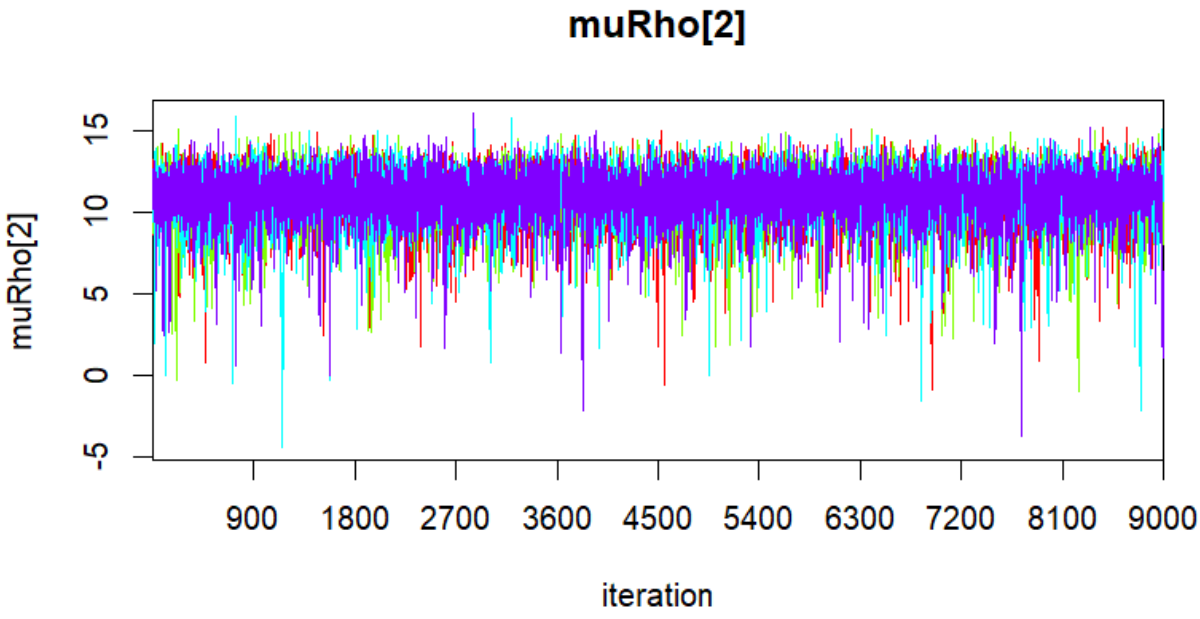

deltaRho
deltaRho
20
10
0
-20
900
1800
2700
3600
4500
5400
6300
7200
8100
9000
iteration


muBeta[1]
muBeta[1]
15
10
5
0
-5
900
1800
2700
3600
4500
5400
6300
7200
8100
9000
iteration


muBeta[2]
muBeta[2]
0
1
2
3
4
5
6
900
1800
2700
3600
4500
5400
6300
7200
8100
9000
iteration


muRho[1]
muRho[1]
-10
0
10
20
30
900
1800
2700
3600
4500
5400
6300
7200
8100
9000
iteration

[redacted]


muRho[2]
muRho[2]
15
10
5
0
-5
900
1800
2700
3600
4500
5400
6300
7200
8100
9000
iteration