METHOD ARTICLE

REVISED Enhancing experimental design through Bayes factor design analysis: insights from multi-armed bandit tasks

[version 2; peer review: 1 approved, 3 approved with reservations]

Sarah Schreiber[1], Danielle Hewitt[2], Ben Seymour[1,2], Wako Yoshida[2,3]

[1]Institute of Biomedical Engineering, University of Oxford, Oxford, England, OX37DQ, UK
[2]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, England, OX39DU, UK
[3]Department of Neural Computation for Decision-making, Advanced Telecommunications Research Institute International, Kyoto, Japan

v2 **First published:** 01 Aug 2024, **9**:423
https://doi.org/10.12688/wellcomeopenres.22288.1
Latest published: 07 May 2026, **9**:423
https://doi.org/10.12688/wellcomeopenres.22288.2

Abstract

Bayesian statistics offers a flexible framework that supports iterative updating of hypotheses and the incorporation of prior information, amongst other advantages. Although well established for retrospective analysis, the application of Bayesian methods to prospective analysis is less well developed, especially when used in combination with computational model-based analysis of behavioural data in cognitive neuroscience. It is therefore important to establish effective methods for testing and optimising experimental designs for these purposes. One framework for a prospective approach is Bayes factor design analysis (BFDA), which can be used alongside latent variable modelling to evaluate and visualise the distribution of Bayes factors for a given experimental design. This paper provides a tutorial-style analysis combining BFDA with latent variable modelling to evaluate exploration-exploitation trade-offs in the binary multi-armed bandit task (MAB). This is a complex example of human decision-making with which to investigate the feasibility of differentiating latent variables between groups as a function of different design parameters. We examined how sample size, number of games per participant and effect size affect the strength of evidence supporting a difference in means between two groups. To further assess how these parameters affect experimental results, metrics of error were evaluated. Using simulations, we demonstrated how BFDA can be combined with latent variable modelling to evaluate and optimise parameter estimation of exploration in the MAB task, allowing inference of the mean degree of random exploration in a population, as well as between groups. However, BFDA indicated that, even with large samples and effect sizes, there may be some circumstances

Open Peer Review

Approval Status ? ? ✓ ?

	1	2	3	4
version 2 (revision) 07 May 2026			✓ view	? view
version 1 01 Aug 2024	? view	? view	? view	

1. **Brandon S Coventry**, University of Wisconsin-Madison, Madison, USA
2. **Dominik Bach**, University of Bonn, Bonn, Germany
3. **Jessica Schaaf**, adboud University Donders Institute for Brain Cognition and Behaviour (Ringgold ID: 198328), Nijmegen, The Netherlands
Radboudumc (Ringgold ID: 6034), Nijmegen, The Netherlands
4. **Tommaso Costa**, University of Turin, Turin, Italy

Any reports and responses or comments on the article can be found at the end of the article.

where there is a high likelihood of errors and a low probability of detecting evidence in favour of a difference when comparing random exploration between two groups performing the bandit task. In summary, we show how BFDA can prospectively inform design and power of human behavioural tasks.

Plain Language Summary

Optimising the design of experiments before collecting data is an important part of cost-efficient and responsible research. However, this is challenging as experimental designs and analysis methods get complicated, as there are many parameters and factors that need to be balanced. These include the sample size and the number of trials each participant completes, but also parameters that are more specific to the experiment. In this work, we show how the use of an approach called 'Bayesian Factor Design Analysis' can optimize the experimental design of a human decision-making task, whereby participants choose between two options with varying outcomes. The objective of this task is to compare information-seeking ('exploratory') behaviour between two groups, which is an important problem in cognitive sciences - proposed, for example, as a key mechanism underlying the development of chronic pain. However, before starting clinical studies, we should first identify how easily we can detect differences between patients and healthy controls, so we know how much data we need to collect. We show that this approach effectively allows estimating the average level of information-seeking behaviour in a group, as well as between groups, according to various parameters. The analysis can identify situations with a high probability of error and a low probability of successfully detecting a difference between two groups.

Keywords

design calculation, Bayes factor, design analysis, sample size, statistical evidence, exploration, decision-making

Corresponding author: Ben Seymour (ben.seymour@ndcn.ox.ac.uk)

Author roles: Schreiber S: Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hewitt D**: Writing – Original Draft Preparation, Writing – Review & Editing; **Seymour B**: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; **Yoshida W**: Conceptualization, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The work was funded by Wellcome Trust (214251/Z/18/Z, 203139/Z/16/Z, and 203139/A/16/Z), EPSRC (EP/W03509X/1), IITP (MSIT 2019-0-01371) and JSPS KAKENHI (22H04998). This research was also part supported by the NIHR Oxford Health Biomedical Research Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. This work was supported by a fellowship of the German Academic Exchange Service (DAAD). W.Y. was funded by Versus Arthritis (ARUK-21537) and MRC 713 (MR/W027593/1), UK.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Schreiber S, Hewitt D, Seymour B and Yoshida W. **Enhancing experimental design through Bayes factor design analysis: insights from multi-armed bandit tasks [version 2; peer review: 1 approved, 3 approved with reservations]** Wellcome Open Research 2026, **9**:423 https://doi.org/10.12688/wellcomeopenres.22288.2

First published: 01 Aug 2024, **9**:423 https://doi.org/10.12688/wellcomeopenres.22288.1

REVISED **Amendments from Version 1**

This article has been updated in response to reviewer comments. Several sections of the text have been revised to improve clarity, precision, and consistency, particularly in the description of the task and the statistical methodology. Minor adjustments to figures and documentation were also made to better align with the revised text.

Any further responses from the reviewers can be found at the end of the article

Introduction

Designing an effective experiment is a multifaceted process that begins with formulating clear research questions and hypotheses, selecting appropriate methodologies, and aligning design choices with the underlying theoretical framework. Beyond these conceptual considerations, the appropriate analyses and practical parameters must be carefully determined to ensure valid, reliable, and adequately powered statistical inference (Czitrom, 2012; Garud *et al.*, 2017; Tůmová *et al.*, 2018). One such practical parameter is the sample size, which is commonly determined by conducting power calculations. This process is an example of a prospective design analysis aimed at optimizing study outcomes. Usually, this is followed by a retrospective evaluation of the design after data collection to test the statistical significance of a result. To improve a study's effectiveness, it is recommended to conduct a thorough prospective design analysis rather than relying purely on a retrospective approach (Gelman & Carlin, 2014). Prospective design analyses can help optimize the use of available resources, which has become increasingly important considering recent concerns about 'research waste' (Ioannidis *et al.*, 2014; Macleod *et al.*, 2014; Storz-Pfennig, 2017). These approaches align with efforts to improve reproducibility and replicability, such as preregistration and sharing of code and data (Munafò *et al.*, 2017; Vize *et al.*, 2025). As BFDA requires the primary analysis to be specified a priori, it integrates well with these practices.

Traditionally, the focus of predictive analyses has been on determining the sample size required to ensure adequate statistical power to detect meaningful effects, which is a useful step to ensure the quality and validity of the experiment and all conclusions drawn from it. Methodologies for determining sample sizes have long primarily relied on frequentist statistics, but there has been an ongoing critique among statisticians and methodologists regarding these approaches. Among the main concerns are the common misinterpretation of *p*-values and significance testing (Benjamin *et al.*, 2018; Gigerenzer, 2004; Morrison & Henkel, 2017; Wagenmakers, 2007; Wagenmakers *et al.*, 2018). Concurrently, Bayesian methods are gaining recognition for their advantages, which include the incorporation of prior knowledge into statistical processes, the ability to quantify evidence for both null and alternative hypotheses, accommodate non-normal data, and directly represent uncertainty through probability distributions (Jeffreys, 1935, 1961; Wagenmakers, 2007; Blackwell & Ramamoorthi, 1982; Gelman & Shalizi, 2013; Kruschke, 2010, 2013). These considerations have led to arguments in favour of a more balanced approach between frequentist methods and Bayesian frameworks in statistical methodology. Thus, as Bayesian statistical methods gain popularity, it is crucial to have the necessary tools to perform a comprehensive Bayesian design analysis.

Schönbrodt and Wagenmakers proposed Bayes factor design analysis (BFDA) as a method for design analysis (Schönbrodt & Wagenmakers, 2018). This framework is based on the Bayes factor, a continuous measurement weighing the evidence for one hypothesis over another (Morey & Rouder, 2011). The Bayes factor can be compared to decision thresholds that indicate different strengths of evidence for the null hypothesis as well as the alternative hypothesis respectively (Jeffreys, 1961). BFDA evaluates the distribution of the Bayes factors for a given experimental design, providing a powerful alternative to frequentist a priori power analyses.

BFDA assumes that the variable of interest is defined at the population level, and that observations are sampled from this population. In well-established paradigms, distributional properties and effect size estimates of the variable at population level can be informed by previous research. In novel paradigms, standard distributions (e.g. the normal distribution) may be assumed and sensitivity analyses across possible effect sizes can be used to assess robustness (Schönbrodt & Wagenmakers, 2018). For each sample generated under this population model, the comparative evidence between the null hypothesis and the alternative hypothesis is measured by calculating the Bayes factor. Repeating this process yields a distribution of Bayes factors, which can be used to evaluate and compare design approaches and thus optimise the experimental setup.

In this work, we focus on what Schönbrodt and Wagenmakers call a fixed- *n* design, where each analysed sample is of a fixed sample size.

Previous literature on BFDA presents examples on how this approach can be used to calculate the probability of errors and how the sample size affects the probability of obtaining a Bayes factor of a certain value (Schönbrodt & Wagenmakers, 2018; Stefan *et al.*, 2019, 2024). These examples are based on variables that are directly measurable. However, in

psychology and neuroscience, we often deal with latent variables that first need to be inferred from the data collected. Such cases are commonly addressed using computational models (e.g. of behaviour or neural responses), including hierarchical Bayesian models, which allow latent parameters to be estimated while accounting for variability at multiple levels (Cronin *et al*., 2010; Gelman, 2006; Gelman & Hill, 2006).

Here, we consider an example problem of differentiating different levels of exploratory choices based on learned values, using the application of reinforcement learning models. This is a problem that directly relates to computational models of neurological and psychiatric disease, including chronic pain and depression (Krypotos *et al.*, 2022a). For instance, in the classic 'Fear Avoidance' model of chronic pain, individuals with acute or subacute musculoskeletal pain are proposed to excessively avoid engaging in physical activity as they approach the recovery period, because of a failure to adequately explore movement actions that might no longer be as painful as expected (Vlaeyen *et al.*, 2016). This leads to a cycle of inactivity and physical deconditioning, which itself ultimately worsens pain. However, the hypothesis as to whether this genuinely relates to impaired exploratory behaviour has not been tested, as it ideally requires a model-based analysis of exploratory behaviour, considering various confounding factors.

The goal of this work is a prospective design analysis, specifically to evaluate how sample size, number of games per participant, and effect size influence Bayes factors and parameter recovery for a latent exploration parameter in a multi-armed bandit task. We first analysed the accuracy of Bayesian parameter estimations of latent variables, within a population, as well as between two groups, using simulated behavioural data in a multi-armed bandit task (Krypotos *et al.*, 2024). In a second step, we combined BFDA with simulations of behavioural data to explore the relationship between sample size and the strength of evidence for both null and alternative hypothesis. This included examining the probability of substantial evidence for an incorrect hypothesis and the probability of insufficient evidence for the correct hypothesis. We also considered how other factors and considerations of realistic experimental design affect these properties.

Methods

Bayes factor design analysis

Bayesian statistics. We begin by outlining the central concepts and key formulas of Bayesian statistics. Readers seeking more detailed explanations, formal derivations, and worked examples are referred to Kruschke (2014) and Hudson (2021). Bayesian statistics is a method of combining evidence from observed data with prior information and beliefs (Bayes, 1991; Kruschke, 2015a; Kruschke *et al.*, 2012). This means we are updating our beliefs based on new information in a probabilistic manner, considering the uncertainty of our prior beliefs as well as the collected data. Bayes' theorem describes this idea mathematically, as

$$p(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{p(Y)}.$$

Here θ represents the parameter being estimated, which generates the collected dataset. The resulting posterior distribution $p(\theta|Y)$ approximates the true probability distribution of θ on the basis of our prior beliefs and the available data (Y). The likelihood $p(Y|\theta)$ quantifies the probability of measuring the observed dataset for various values of the unknown parameter we are aiming to estimate. The prior distribution $\pi(\theta)$ represents our knowledge about the distribution of θ ahead of data acquisition. The chosen prior can heavily influence the outcome of an analysis and should therefore be chosen carefully, as it will bias the estimation. When prior knowledge about the underlying distribution is limited, researchers often employ uninformative or weakly informative priors, which can nevertheless reflect plausible bonds for the parameters. Given the subjective nature of priors, their selection should be transparently reported and sufficiently motivated. In addition, sensitivity analyses can be conducted to evaluate the robustness of results with respect to the chosen prior. For further discussions on priors, refer to further literature such as Gelman (2006), Van Dongen (2006), or Stefan (2019).

To calculate the posterior according to (1) we need the marginal density $p(Y)$, which is calculated by

$$p(Y) = \int p(Y|\theta)\,\pi(\theta)d\theta.$$

This results in the posterior

$$p(\theta|Y) = \frac{p(Y|\theta)\pi(\theta)}{\int p(Y|\theta)\pi(\theta)d\theta}.$$

The marginal density p(Y) is therefore a simple normalizing constant, and the posterior depends only on the likelihood $p(Y|\theta)$ and the prior $\pi(\theta)$.

The likelihood $p(Y|\theta)$ can be calculated based on the experimental design. Provided that the collected dataset Y consists of n independent measurements, $Y = [y_1, y_2, ..., y_n]$, the likelihood of measuring this dataset is the product of the likelihoods of measuring each independent data point within the dataset

$$p(\mathrm{Y}|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

The likelihood function for each independent data point is specific to the experimental setup, and the chosen outcome measure.

From the posterior, a point estimate of θ can be calculated using

$$\widehat{\theta} = \int \mathrm{p}(\theta|\mathrm{Y})\theta \mathrm{d}\theta.$$

Bayes factor. Once the likelihoods have been calculated for each group, we can compare the distribution of θ between two groups, for example between patients and healthy controls. A Bayesian approach can be used to estimate this difference. Using the individual likelihoods $p(Y_1|\theta_1)$ and $p(Y_2|\theta_2)$ calculated as previously described, we can compute the probability distribution of the difference in θ between groups, $\Delta\theta = \theta_1 - \theta_2$. Considering two independent continuous random variables X and Y, the probability density function of the difference $Z = X - Y$ can be calculated by convolving the respecting probability density functions $f_Z(z) = \int f_X(x) f_Y(x-z)\, dx$. In our case we can calculate the convolution of the two likelihoods to get the probability density function of the difference.

$$p(Y|\Delta\theta) = \int p(Y_1|\theta_1) p(Y_2|\theta_1 - \Delta\theta) d\theta_1.$$

This probability density function can then be used to obtain a point estimator for the difference in θ between the two groups. If the objective is to determine whether a difference exists, regardless of the exact value, we can calculate the Bayes factor.

$$BF_{10} = \frac{m_1(Y)}{m_0(Y)},$$

with $m_1(Y)$ denoting the marginal density of the alternate hypothesis based on data Y and $m_0(Y)$ describing the marginal density of the null hypothesis. The Bayes factor is a mathematical description of Bayesian hypothesis testing (Jeffreys, 1935, 1961; Johnson *et al.*, 2023; Kruschke, 2015b). It weighs the evidence for one hypothesis against another. In our case, we are comparing the hypothesis that there is a difference in θ between two groups, H_1: $\Delta\theta \neq 0$, against the null hypothesis, H_0: $\Delta\theta = 0$. Thus, our Bayes factor is calculated as:

$$\mathrm{BF}_{10} = \frac{\int \mathrm{p}(\mathrm{Y}|\Delta\theta)\pi(\Delta\theta)\mathrm{d}\Delta\theta}{\mathrm{p}(\mathrm{Y}|\Delta\theta = 0)}.$$

A visual representation of the Bayes factor is shown in Figure 1a. An advantage of the Bayes factor is that it is a continuous measurement on the evidence for one hypothesis over another. A higher factor corresponds to stronger evidence in favour of our alternative hypothesis, just as a smaller factor of less than one corresponds to stronger evidence in favour of the null hypothesis. A widely used scale for categorising the Bayes factor is based on work by Jeffreys (Jeffreys, 1935, 1961; Kass & Raftery, 1995) and is shown in Figure 1b.

Evaluation measurements. To give some examples of the benefits of using modelling in the design stages of an experiment, behavioural data from a multi-armed bandit task was simulated. An analysis was conducted to investigate the impact of sample size, number of games per participant, and difference in population mean on the computed Bayes factors. When considering the Bayes factor, it is important to note that it can have a significant variance (Pfister, 2021). This means that the same experimental design and analysis will produce Bayes factors of different strengths of evidence when replicated. This is the fundamental principle of BFDA. It involves repeated sampling and analysis of a simulated population followed by an analysis of the distribution of Bayes factors across these samples (Schönbrodt & Wagenmakers, 2018). To compare distributions for different experimental conditions, we can examine the frequencies of Bayes factors exceeding common thresholds, as outlined in Figure 1b. We can achieve this by running a set of simulations and

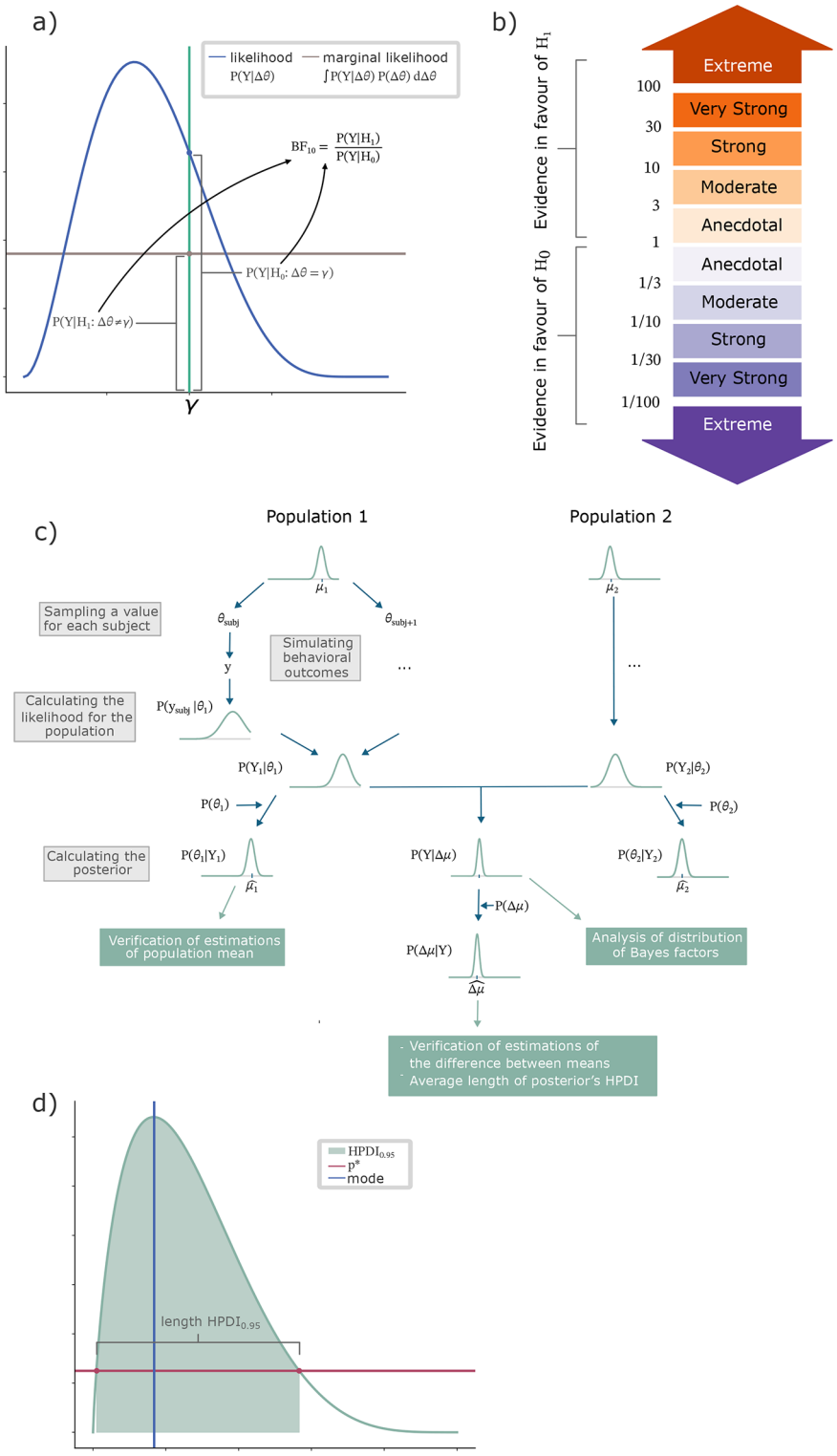

Figure 1. Summary of methods. BFDA was combined with latent variable modelling for an example of a MAB task. (**a**) The Bayes factor was used as a measure of evidence in favour of a difference in mean. (b) Decision thresholds for the Bayes factor indicating strengths of evidence for either hypothesis (Jeffreys, 1935, 1961; Kass & Raftery, 1995). (c) Repeated samples were taken from two populations and used to simulate behavioural outcomes of individual subjects. Based on these simulations the difference in means between the two groups was analysed by examining the distribution of Bayes factors, as well as probabilities of errors in the estimations of the difference. (d) Example of a 95% HPDI.

calculating the average number of simulations with a Bayes factor indicating each of these strengths of evidence (Figure 1c). For computational purposes, the parameter space was discretized over a bounded grid, and all integrals with respect to parameters were evaluated as finite sums over this grid.

A further analysis evaluates the accuracy of the estimation of the difference between two groups by assessing the estimation error and the average length of the highest probability density interval (HPDI) to aid in determining an optimal sample size. The HPDI is a type of credible interval, representing a range of values within which the parameter has a certain probability of falling (e.g. 95%). Among all possible intervals, the HPDI is the credible interval that includes the highest probability densities. Therefore, a 95% HDPI would include the range of values with the highest densities and 95% probability. The interval is defined as:

$$C_{1-\alpha} = \{\theta : p(\theta|Y) > p^*\}$$

with

$$\int_{C_{1-\alpha}} p(\theta|Y)d\theta = 1 - \alpha.$$

An example of a HPDI is shown in Figure 1d.

The length of the HPDI can serve as a measurement of uncertainty, with narrower intervals reflecting more precise estimates.

The average length of the posterior's *HPDI* can be calculated in a two-step approach (Joseph & Bélisle, 1997). Let l' be the length of the $HPD_{1-\alpha}$ interval for a given dataset Y_i. The average length l^* can be calculated from l' by multiplying it by the probability of Y_i being the outcome for this measurement and integrating over all possible outcomes $\gamma = [Y_1, Y_2, ...]$

$$l^* = \int l'(Y)p(Y|n)dY.$$

$p(Y|n)$ is the posterior predictive density and can be computed as

$$p(Y|n) = \int p(Y|\theta,n)\pi(\theta)d\theta.$$

Both the length of the HPDI and the predictive posterior are functions of the sample size of each group n. A higher n leads to a smaller HPDI, which indicates a higher certainty about the value of the estimated parameter.

In addition to serving as a tool for monitoring the quality of the estimation algorithm, the average length can also be used as a method for determining the sample size by predefining a maximum length. This is the basis of the average length criterion (ALC) (Joseph & Bélisle, 1997). The sample size is determined as the smallest n, for which the average length of the HPDI is smaller than a set maximum length l_{max}.

$$\int l'(Y)p(Y|n)dY \leq l_{max}.$$

Other common methods for sample size determination include the average coverage criterion, and the worst outcome criterion (Cao *et al.*, 2009).

The exploration-exploitation dilemma

Multi-armed bandit task. To validate the analysis pipelines and to provide a concrete example for the proposed approach, a binary two-armed bandit task was simulated. The MAB is a widely-used paradigm to investigate the exploration-exploitation dilemma in behavioural science (Danwitz *et al.*, 2022; Daw *et al.*, 2006; Gershman, 2019), in which the agent (for example the participant or a reinforcement learning agent) repeatedly chooses between multiple actions or "arms". Over time the agent learns the reward probabilities associated with each arm. To maximise the cumulative outcome across trials, the agent must balance exploring the available arms in order to reduce uncertainty about their outcomes (exploration) and using the information they have gathered so far to choose the optimal option (exploitation).

The outcomes can be continuous, for example in the form of a scalar monetary reward, where participants learn the reward distribution (e.g. mean and uncertainty), or binary, such as the presence versus absence of a food reward or a negative (aversive) stimulus. The binary version of the MAB task is often used in pain research, where a painful stimulus (e.g., an

electrocutaneous shock) is administered. The two possible outcomes would be the delivery of an aversive (e.g. painful) stimulus and its absence (Krypotos *et al.*, 2022a, 2022b). In the computational modelling of these outcomes, each outcome must be assigned a numerical value. Here, the outcome of each choice was coded as 0 or −1, where −1 represents an aversive outcome.

Explore-exploit trade-off in the Horizon task. To facilitate differentiation between exploration and exploitation, the agent has information on both options available to them prior to their first choice, so that the agent possesses information to base their exploitation on (Wilson *et al.*, 2014). This is achieved by presenting the agent with four actions and their immediate results, after which they are free to make their own choices. The task was implemented with a fixed horizon of 10 choices, defining the number of trials within each game, in that each game consisted of four observed trials followed by 6 free choice trials. The outcome probabilities were combinations of the probabilities 0.1, 0.3, and 0.9, with the probability assigned to each arm changing after each game.

Calculating the likelihood of a certain outcome in this case is somewhat complex, as we must model human behaviour. If we want to infer a participant's tendency towards exploration as opposed to exploitation from the recorded data, we must make several assumptions about their behaviour. Reinforcement learning provides a straightforward approach to modelling this type of decision-making.

Calculation of the likelihood. The binary two-armed bandit task was simulated as described and the number of draws with an aversive outcome was recorded. The probability of receiving a certain number of aversive outcomes in one game is dependent on the choices of the participant on one hand and the reward probabilities of each arm on the other hand. The latter are set in the experimental setup and thereby known. The former was modelled by a softmax algorithm

$$\mathrm{p}\,(\mathrm{aj}|\tau) = \exp\,(\mathrm{Q}\,(\mathrm{aj})\tau)\sum \mathrm{i} = 1\mathrm{kexp}(\mathrm{Q}\,(\mathrm{ai})\tau), \tag{14}$$

where the parameter τ determines the degree of exploration. A higher value of τ correlates to a higher degree of exploration. The Q-value is a weighted average of previous rewards

$$\mathrm{Qnj} = \mathrm{Qn} - 1\mathrm{j} + \alpha\,(\mathrm{Rn} - 1 - \mathrm{Qn} - 1\mathrm{j}), \tag{15}$$

with R_i representing the reward on trial *i*.

In each trial, the agent chooses between two arms. The probabilities of choosing the respective arms are dependent on the previous choices and received rewards. Once a choice is made, there is a certain probability of receiving an aversive outcome.

For each choice, there are four possible combinations of arm chosen and aversive or neutral reward. Because the decisions depend on a trial-by-trial update of the Q-values, we need to consider not only the number of aversive outcomes that were previously received when playing a certain arm, but also the order of these outcomes. For six free choice trials there are a total of 4^6 possible paths, or combinations of choices and outcomes. We can calculate the probability of each of these paths as a product of the probabilities of the choices and corresponding outcomes within this path. The probabilities for each choice can be calculated from (14).

As a performance metric, we recorded the total number of aversive outcomes received over the course of one game. In our case, the agent has six free choices per game, therefore, they can receive between zero and six aversive outcomes. To obtain the probability of each possible value of the performance metric, i.e. the number of aversive outcomes n_a in a game, we iterate through all possible paths resulting in this outcome and sum up their likelihoods. As we have seen in (14), this probability is dependent on the variable τ, which determines the degree of exploration in the behaviour. The iterations and calculations can be repeated for all possible outcomes n_a, as well as a set of different τ, which will give us $p(y_i|\tau)$. The likelihood for a collected dataset, $p(Y|\tau)$, can then be calculated according to (4).

To summarise, the primary outcome per game is the number of aversive outcomes (0–6), while the main inferred parameter is the exploration parameter τ. The between group comparison focusses on the difference in the exploration parameter $\Delta\mu_\tau$. For both tau and the difference in population means ($\Delta\mu_\tau$), we used a discrete uniform prior defined over a bounded parameter space ($\tau \in [0.01, 3]$ and $\Delta\mu_\tau \in [-3, 3]$).

In the present simulations, the learning rate α was arbitrarily fixed at 0.1 for illustrative purposes. Although learning rates may vary, particularly in aversive contexts (Krypotos *et al.*, 2022b; Wang *et al.*, 2018), our conclusions are not contingent on this specific choice. The Q-values for both arms were initialised with $Q_0 = 0$ and updated after each choice.

Results

The use of the BFDA as a method for prospective design analysis allows us to systematically evaluate design parameters, including sample size, number of games, task-specific features such as number of arms and aversive outcome probabilities, and the planned analysis approach. In our example experiment of the binary MAB with aversive outcomes, our hypothetical question is to compare the exploration parameter τ between two groups. This is a latent parameter which influences the agents' choices (eq.14). Our analysis consists of three main steps summarized in Figure 1c. First, we verify the estimations of population mean for one population. Next, we compare the population means between two group by examining the distribution of Bayes factors, which indicate whether there is evidence for or against a difference in population mean between groups. This analysis is carried out to determine possible effect sizes and the effect of experimental parameters, such as the number of participants, as well as the number of games per participant on the evidence. In the last step, we analyse potential errors when estimating a difference in means between groups and how they are affected by the experimental parameters (number of participants and number of games per participant).

Validation of the estimation algorithm

To validate the estimation algorithm, the agent's choices on the multi-armed bandit task were simulated for a range of exploration parameters. Each simulated dataset consisted of a range of sample sizes between 10 and 58 in increments of 2 with $n_{\text{games}} = 150$ games per 'participant'. For each simulated participant, the individual exploration parameter τ, as defined in the softmax algorithm (14), was drawn from a normal distribution with population mean μ_τ ranging from 0.05 to 1 in increments of 0.05, and a standard deviation of $\sigma = 0.02$. The goal was to then estimate the population mean within the Bayesian framework as described (see Figure 1c, Validation of estimations of population mean).

To evaluate the accuracy of this estimation, we compared the point estimates calculated as stated in (5) with the true exploration parameters. This was done using a linear regression model with point estimates as the dependent variable and true mean difference, sample size, and their interaction as predictors. The regression model, as well as the subsequent models, were estimated using ordinary least squares in the seaborn python package (Seabold & Perktold, 2010). These regression analyses and correlations are descriptive summaries of the observed relationships and are not part of the Bayesian inferential framework or the Bayes factor-based design analysis. They are used solely as a descriptive check to assess whether the simulated data contain informative signal regarding the accuracy of the estimated population means.

We then calculated the mean squared estimation error (MSEE) and evaluated its relationship with sample size using Spearman's rank-order correlation. We also computed a regression model with the MSEE as the dependent variable and the sample size, true population mean, and their interaction as predictors, allowing for nonlinear relationships. After inspection, the sample size was logarithmically transformed, and the population mean was exponentially transformed. The scaling constants used in these transformations were determined in prior curve-fitting procedures.

Multiple regression analysis was used to investigate the correlation between the mean estimated exploration parameter with the sample size and the true population mean. The overall fit of the regression model was statistically significant ($F(3,496) = 136300$, $p < 0.001$, $R^2 = 0.999$). The sample size and population mean, as well as their interaction were significant predictors of the mean estimations ($t = 7.246$, $p < 0.001$; $t = 265.589$, $p < 0.001$; $t = -17.442$, $p < 0.001$). The mean squared estimation error (MSEE) revealed strong negative correlations with the sample size across all simulated mean exploration parameters ($r(23) < -0.96$, $p < 0.001$). The relationship between the MSEE and the sample size, as well as population mean, was analysed using multiple regression analysis. The overall fit of the regression model was statistically significant ($F(3,496) = 7763$, $p < 0.001$, $R^2 = 0.979$). The transformed sample size and mean, as well as their interaction, were significant predictors of the MSEE ($t = 9.284$, $p < 0.001$; $t = 126.156$, $p < 0.001$; $t = -96.491$, $p < 0.001$). These results demonstrate that the algorithm is successful in estimating the mean exploration parameter within a group. It is also shown that the MSEE decreases with increasing sample size, which indicates that the estimations increase in accuracy with more participants. This is a crucial point that must be validated before proceeding with the analysis, as it forms the foundation for subsequent considerations, such as sample size analyses.

Next, we analysed the estimated difference in mean between the two populations by comparing the estimated difference between the two populations $\widehat{\Delta\mu_\tau}$ to the true difference $\Delta\mu_\tau$. 250 simulations were run for each combination of $\Delta\mu_\tau$ ranging from 0 to 1 in increments of 0.05 and sample size per group ranging from 10 to 58 in increments of 2. A linear regression model with point estimates of the difference as the dependent variable and true mean difference, sample size, and their interaction as predictors was used as a descriptive summary of the correspondence between estimated and generative values.

Again, these analyses are not part of Bayesian inferential framework. The MSEE was calculated for the estimated difference and its relationship with sample size and mean difference was examined using a regression model with MSEE

as the dependent variable and the sample size, true mean difference, and their interaction as predictors, incorporating nonlinear transformations of the predictors. Log transformation was applied to the sample size, and exponential transformation was applied to the difference in population mean, with scaling constants established in earlier curve-fitting operations.

Multiple regression analysis was used to assess the relation between the estimated difference in means and the sample size, as well as the true difference in population means. The overall fit of the regression model was statistically significant ($F(3,496) = 137900$, $p < 0.001$, $R^2 = 0.999$). The difference in population means as well as the interaction between the difference in means and the sample size were significant predictors of the estimated difference in means ($t = 266.473$, $p < 0.001$; $t = -16.932$, $p < 0.001$), while the sample size was not ($t = 0.522$, $p = 0.602$). The MSEE was calculated for the simulated range of *n*. The relationship between the MSEE and the sample size, as well as the difference in means, was explored using multiple regression analysis. The overall fit of the regression model was statistically significant ($F(3,496) = 3506$, $p < 0.001$, $R^2 = 0.955$). The transformed sample size and difference in means, as well as their interaction, were significant predictors of MSEE ($t = -24.571$, $p < 0.001$; $t = 54.985$, $p < 0.001$; $t = -42.632$, $p < 0.001$). These results indicated that the algorithm is effective in estimating the difference in the random exploration parameter between two groups. It is also highlighted that the MSEE decreases with increasing sample size, indicating that the estimations become more precise with an increased number of participants. Having validated that the analysis can accurately estimate the population mean, as well as the difference in means between two groups, we can use it in combination with BFDA.

Bayes factor design analysis

BFDA entails simulating behavioural data from repeated samples of a population. The analysis of each simulated sample follows the same procedure as the planned analysis of the real dataset. The experiment's design can then be evaluated by calculating the probability of the analysis supporting the null or alternative hypothesis, as well as the probabilities of errors, such as finding evidence in favour of the wrong hypothesis or overestimating the effect size. This analysis looks at the Bayes factor as a function of the difference in means $\Delta\mu_\tau$, sample size n of each group and number of games per participant n_{games}. In addition, the interplay between n and n_{games} was further investigated by considering the average length criterion for sample size determination.

Determining possible effect sizes. To evaluate the effects of sample size and difference in means on the evidence for a difference between groups, we performed the BFDA. First, we drew random samples of the exploration parameter τ from the two populations with different means, with each sampled value representing one simulated participant. Behavioural outcomes within the binary MAB were then generated based on τ, and the distribution $p(Y|\Delta\mu_\tau)$ was computed as described in Equation 6, from which the Bayes factor was calculated (Equation 8). This procedure was repeated 250 times for each combination of the sample size of each group ranging from 10 to 58 in increments of 2 and difference in means ranging from 0 to 0.95 in increments of 0.05. To determine the population means for a given mean difference, the mean of the first population was randomly sampled from a uniform distribution between 0 and $1 - \Delta\mu_\tau$, and the mean of the second population was defined by adding the specified difference. The Bayes factor was calculated for each of the 250 simulations, yielding a distribution of Bayes factors reflecting the evidence for or against a difference in means at each combination of sample size and mean difference. The relative frequencies of the Bayes factor BF_{10} indicating different strengths of evidence is shown in Figure 2 for representative values of $\Delta\mu_\tau$ and n.

$\Delta\mu_\tau$ and n the Bayes factors are more likely to support the correct hypothesis and the stronger the evidence in favour of the correct hypothesis. When the difference in means is zero, the evidence correctly supports the null hypothesis. However, even very small deviations from zero technically contradict the null hypothesis, which can challenge the algorithm. Detecting and strongly supporting such minimal differences requires very large sample sizes.

The relative frequency of a Bayes factor higher than 10 is shown in Figure 3a for the simulated values of $\Delta\mu_\tau$ and n. This Bayes factor would indicate at least strong evidence for the alternative hypothesis, which for this example states that there is a difference in the exploration parameter between two groups. For a true difference in the mean degree of exploration $\Delta\mu_\tau \neq 0$, this probability of at least strong evidence supporting the alternative hypothesis would correspond to the power in frequentist statistics, assuming we reject the null hypothesis for a Bayes factor higher than 10.

Similarly, we can calculate the probability of receiving a Bayes factor BF_{10} smaller than 1/3 which would indicate at least moderate evidence in favour of the null hypothesis (Figure 3b). When this occurs despite a true nonzero mean difference, it constitutes a Type S (sign) error in the Bayesian framework. One way to visualize these probabilities is via a contour plot as shown in Figure 3. Figures 3a,b demonstrate the ability of the algorithm to distinguish a difference in the degree of exploration between two groups as a function of the actual difference and their sample size. As the sample size and true

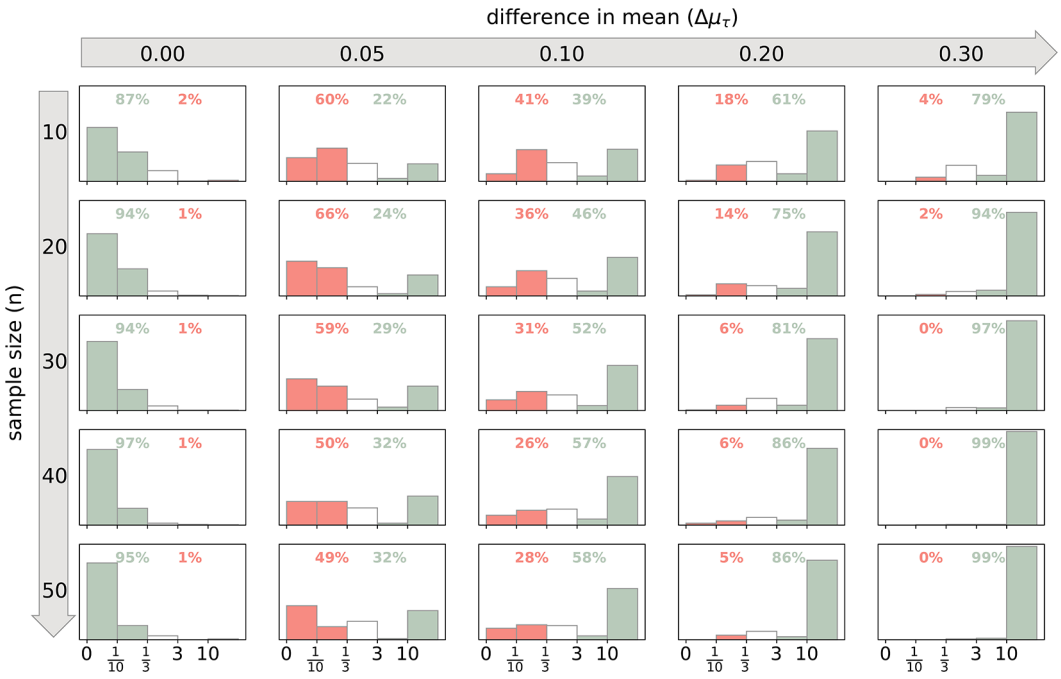

Figure 2. Probabilities of Bayes factors indicating different strengths of evidence. Data was simulated to test for a difference in the degree of exploration between two groups. For each combination of sample size of each group n and the difference in mean $\Delta\mu_\tau$ 250 simulations were performed. The number of games per participant was n_{games} = 150. The green bars show the relative frequency of a Bayes factor indicating at least moderate evidence in favour of the correct hypothesis, while the red bars show the relative frequency of at least moderate evidence in favour of the incorrect hypothesis. The cumulative probabilities for these instances are given as percentages. The white bars show the relative frequencies of anecdotal evidence for either hypothesis.

difference increase, the probability of finding strong evidence for the alternative hypothesis increases. These plots can help to determine the necessary difference in means to support the alternative hypothesis for a given sample size, or vice versa. If an approximation of the true difference in means is known, it is possible to estimate the sample size needed to demonstrate a difference in population means. However, this does not guarantee an accurate estimation of the difference but rather provides evidence for or against the null hypothesis.

Using a rough estimate of $\mu_\tau \approx 0.1$, we can infer from Figure 3a that about 40 participants per group would be needed to achieve an above 50% chance of the analysis yielding Bayes factors above 10, highlighted with a white circle. Adding more participants does not appear to have a substantial impact and only marginally improves the chances. For this exemplary sample size of $n = 40$ and an exploration parameter of $\mu_\tau = 0.1$, the probability of obtaining a Bayes factor smaller than 13 is around 26%. This means that there is a 26% likelihood of finding at least moderate evidence in favour of an incorrect hypothesis. On the other hand, if the null hypothesis were true ($\mu_\tau = 0$), we would expect a Bayes factor greater than 10 with a probability of 0.4%, indicating at least strong evidence in favour of the alternative hypothesis, and a Bayes factor less than 13 with a probability of 96.8%, indicating at least moderate evidence in favour of the null hypothesis. The exact percentages reported here are obtained directly from our simulations, while the broader probability regions are illustrated in the contour plot (Figure 3).

Balancing the number of games per participant and sample size. When designing experiments, it is worth considering the balance between the number of participants and the number of trials each participant completes. The relative frequency of a Bayes factor higher than 10, $Pr(BF_{10} > 10)$, was calculated across 500 simulations per combination of the sample size of each group n and the number of games per participant n_{games}, as shown in Figure 3c. These calculations were carried out for a true difference in mean of $\Delta\,\mu_\tau = 0.3$. This value was chosen arbitrarily to demonstrate how the number of participants and the number of trials affect the evidence in favour of a difference in means. The simulations showed an increase in the relative frequency of a Bayes factor BF_{10} higher than 10 with an increase in the sample size and number of individual trials. This is supported by positive correlations between $Pr(BF_{10} > 10)$ with n and n_{games} (r_s (23) = 0.74, $p < 0.001$; r_s (7) = 0.62, $p < 0.001$) and negative correlations between $Pr(BF_{10} < 1/10)$ with n and n_{games} (r_s (23) = −0.49, $p < 0.001$, r_s (7) = −0.31, $p < 0.001$).

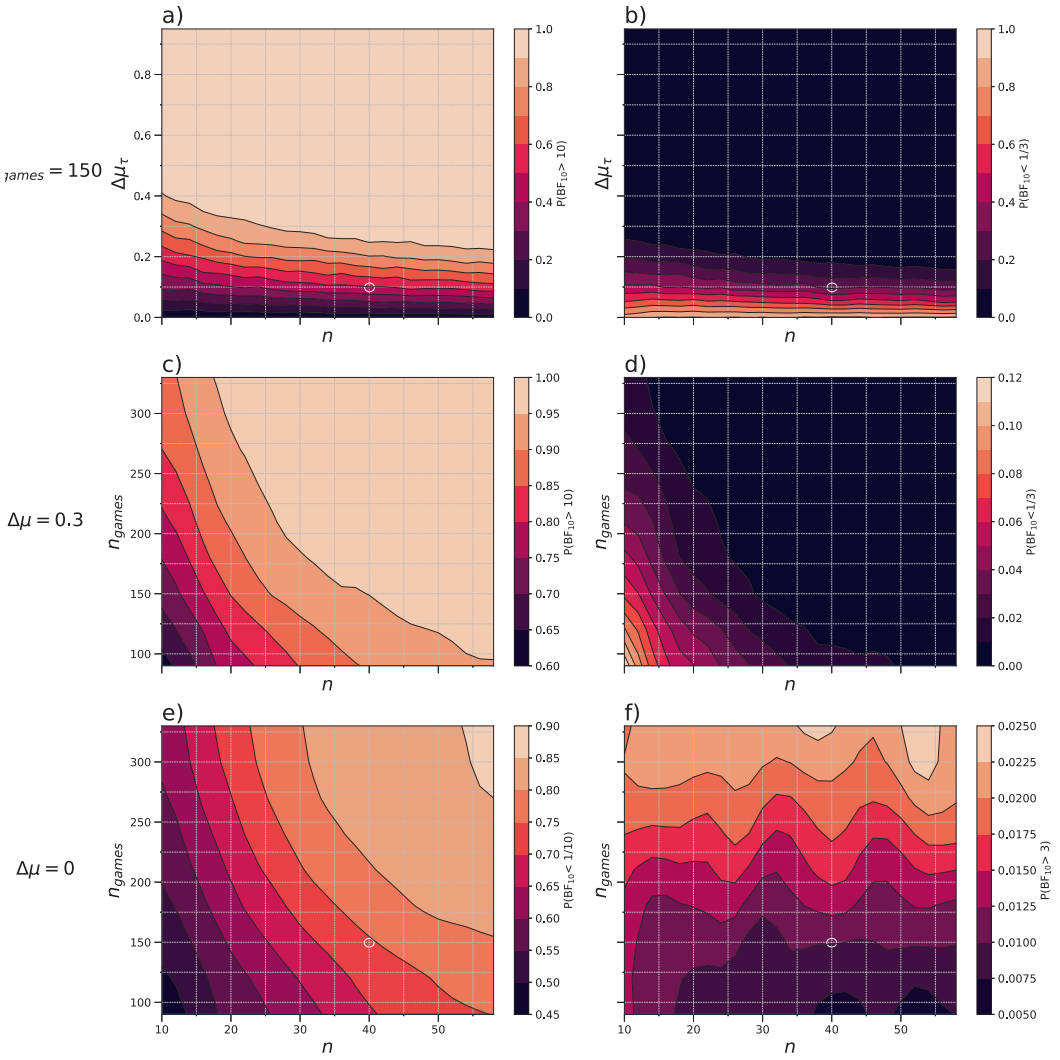

Figure 3. Probabilities of a Bayes factor indicating evidence in favour of the null and alternative hypothesis. Data was simulated to test for a difference in the degree of exploration between two groups. The contour plots display interpolations of the relative frequency of a Bayes factor indicating at least strong evidence in favour of the alternative hypothesis (**a**, **c**, **f**) and of the null hypothesis (**b**, **d**, **e**). The relative frequency was considered as a function of n and $\Delta\mu_\tau$ for $n_{games} = 150$ (**a**, **b**), and as a function of n and n_{games} for the true differences of $\Delta\mu_\tau = 0.3$ (**c**, **d**) and $\Delta\mu_\tau = 0$ (**e**, **f**). The examples discussed in the main text are marked with white and green circles. The contour plots were smoothed with a Gaussian filter (*a,b*: $\sigma = 0.5$; c, d: $\sigma = 0.8$; e, f: $\sigma = 1.3$).

The appropriate balance between n_{games} and n is dependent to external factors. For instance, to achieve a probability of at least 90% for obtaining a Bayes factor greater than 10, we could consider a sample size of 20 participants per group each completing 200 games. Alternatively, a sample size of 30, with each participant completing 130 games could be implemented (Figure 3c, green circles), resulting in a smaller total amount of games played. For both options the probability of wrongfully supporting the null hypothesis is below 2% (Figure 3d, green circles). Assuming all participants complete the same number of trials per hour, the second option would result in less expenses for participants paid at an hourly rate. This would be suitable for an online study, with no additional costs per participant. However, if the study incurs additional costs per participant, the first option may be preferable. When selecting an appropriate sample size, it is important to also ensure that the chosen parameters provide Bayes factors to favour the null hypothesis if it is true. This can be conducted by repeating the above simulations for a true difference of $\Delta\,\mu_\tau = 0$ as shown in Figures 3e,f.

To gain an understanding of how the Bayes factor relates to the true estimate of the difference in mean, we can determine the magnitude error for those simulations that result in a Bayes factor greater than 10. The magnitude error (Type M error) is calculated by dividing the difference between the estimated value and the true difference in mean by the true difference

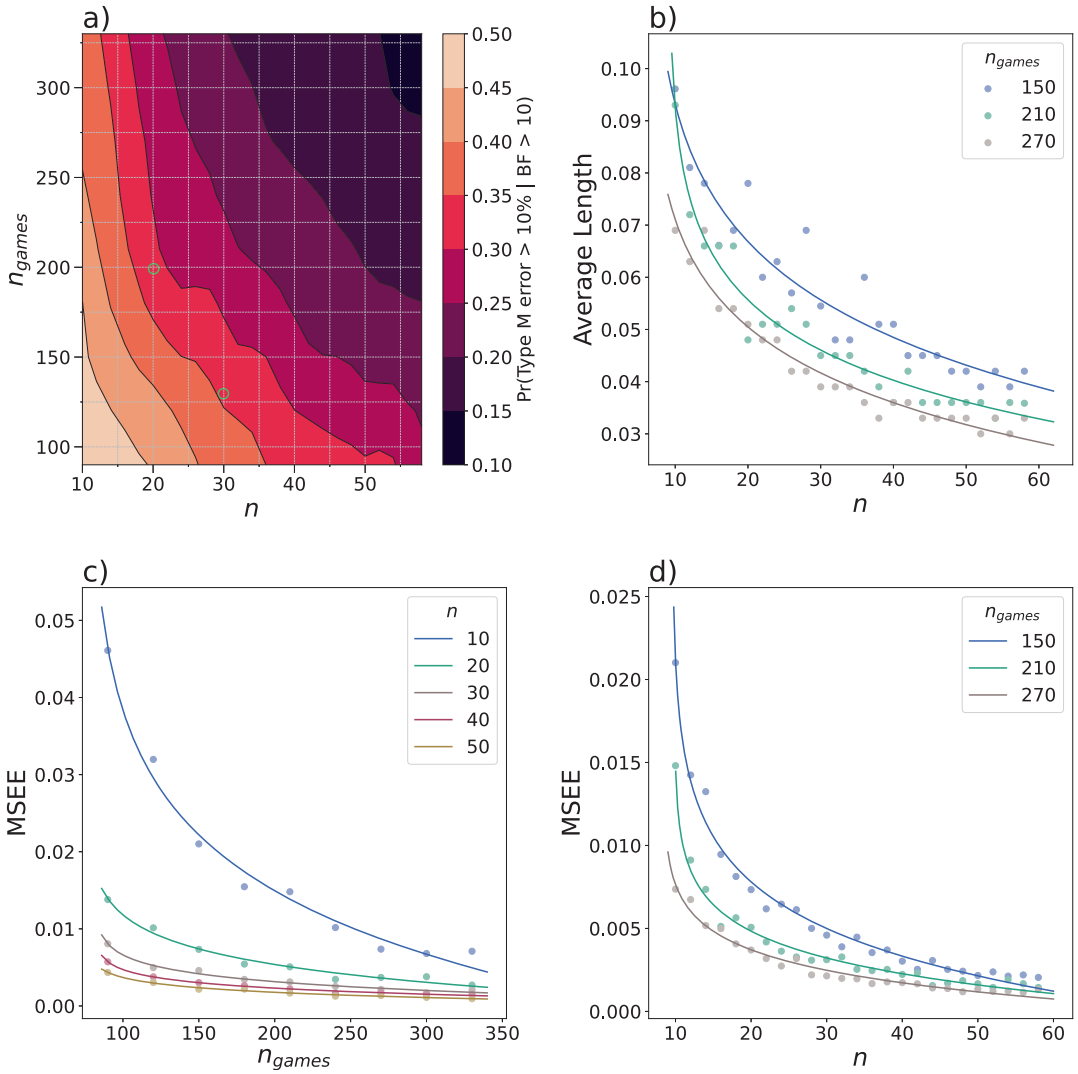

Figure 4. Further considerations for an optimal balance between number of games per participant and sample size. Data was simulated to test for a difference in the degree of exploration between two groups with a true difference of $\Delta\mu_\tau = 0.3$. For each combination of sample size of each group n and the number of games per participant n_{games}, 500 simulations were performed. (**a**) The magnitude error was calculated for all simulations resulting in a Bayes factor greater than 10. The contour plot displays an interpolation of the relative frequencies of a magnitude error exceeding 10%, smoothed with a Gaussian filter ($\sigma = 0.9$). The examples discussed in the main text are marked with green circles. (**b**) The 95% HPDI of the posterior was determined as a function of n and n_{games}, and the average length was calculated. The mean squared estimation errors were calculated and sample values are shown to illustrate the relationship between the mean squared estimation error, n_{games} and n (**c**, **d**).

in mean. This reflects an overestimation of the true effect size in statistically significant results (Gelman & Carlin, 2014). The probability of a magnitude error exceeding 10% decreases with an increase of n and n_{games} (Figure 4a). This may seem trivial, as it is expected that our calculations become more accurate when more data is available. However, the probability of overestimating the effect is an important aspect to consider in both prospective and retrospective design analysis. Simulations as carried out here can aid in choosing optimal values of n and n_{games} to minimize the probability of errors.

To illustrate further, we can consider our example from the previous section, where we compared $n = 30$ with $n_{\text{games}} = 130$ to $n = 20$ with $n_{\text{games}} = 200$. The probability of obtaining a Bayes factor greater than 10 was 90% for both options. To decide between the two options, we can consider the magnitude error. The first option has close to a 35% probability of overestimating the effect by at least 10% for all significant results. For the latter, this probability is between 30% and 35% (Figure 4a, green circles). Therefore, the latter option is less likely to overestimate the true effect size, although these probabilities are still relatively high.

As another measure of the accuracy of our estimation of the effect size, the average length of the $HPDI_{.95}$ was calculated from the simulated data for each combination of sample size ranging from 10 to 58 and n_{games} ranging from 90 to 330 of which representative values are shown in Figure 4b. The average length decreased with increasing n and n_{games}, which corresponds to a narrower posterior distribution and therefore a higher certainty about the value of the estimated parameter. However, the average length does not provide any information regarding the location of the highest probability densities. Consequently, our simulations may result in a highly narrow posterior distribution that is centred around a wrong estimation. To evaluate the validity of our estimations, the MSEE was calculated, which showed a decrease as n and n_{games} increase (Figure 4c,d). If both the error and average length of the HPDI are decreasing with n, the estimations are increasingly accurate.

Discussion

Bayesian statistics offer a robust framework for parameter estimation and hypothesis testing across a range of problems (Ashby, 2006; van de Schoot *et al.*, 2017), and are therefore a valuable alternative to frequentist methods. However, in Bayesian statistics, prospective design analyses are less common compared to their frequentist counterpart, and protocols for effective experimental designs and testing methods need to be established. In this work, we used BFDA in the example of a binary multi-armed bandit (MAB) task with aversive outcomes to analyse the effect of sample size, number of games per participant, and effect size on the probability of obtaining significant evidence to support a null or alternative hypothesis, and the probabilities of incorrectly supporting either hypothesis. This furthers previous work on BFDA (Schönbrodt & Wagenmakers, 2018; Stefan *et al.*, 2019, 2024) by incorporating additional aspects of experimental design which are critical in behavioural sciences. Additionally, we contribute to existing knowledge by integrating latent variable analysis into BFDA, as demonstrated by a practical example using a multi-armed bandit task. The prospective design analysis was expanded by examining the average length criterion for studies investigating precise estimates of the difference between groups.

The example of a multi-armed bandit (MAB) task with prior information was used to examine latent variables in relation to BFDA. Specifically, we considered a hypothetical experiment designed to assess differences in the exploration parameter (τ), estimated via the softmax decision-rule, between two groups. We showed how sample size, number of games per participant and effect size influence the probability of obtaining evidence for group differences in τ. Previous studies have focused on inferring the degree of random exploration from behavioural outcomes in this task (Mizell *et al.*, 2024; Somerville *et al.*, 2017; Waltz *et al.*, 2020; Wilson *et al.*, 2014, 2021), with significant differences in random exploration being identified between certain populations (Mizell *et al.*, 2024) and not others (Mizell *et al.*, 2024; Waltz *et al.*, 2020). These studies employed frequentist methods, which do not provide evidence in favour of a null hypothesis. Therefore, there is no evidence supporting the absence of a difference between the groups. Our study validated a Bayesian analysis for the MAB with prior information using Bayes factors and estimations of the difference. We found that it can be challenging to determine the difference in means between two groups. Large sample sizes are needed for a strong probability of detecting evidence in favour of difference, where present. If there is strong evidence supporting the alternative hypothesis, the probability of overestimating the difference in means is relatively high. In the example we used an uninformative prior, but depending on the experiment and existing literature, this should be adjusted. A well-informed prior can lead to higher probabilities of detecting a difference between groups.

Optimizing experimental parameters can increase the likelihood of finding evidence supporting either null or alternative hypothesis while reducing the likelihood of overestimating the true effect. Previous research has investigated the most efficient design for MAB tasks by optimising a utility function (Valentin *et al.*, 2024; Zhang & Lee, 2010). This function aims to maximise the information gain of each design (Ryan *et al.*, 2016). However, it is important to consider the probability of errors, such as overestimating effects, as well as resource considerations, such as cost and space, as described in our work.

We demonstrated that that the probability of a Bayes factor indicating strong evidence highly depends on the number of games each participant completes. The analyses indicate that, in certain situations, it may be more efficient to increase the number of trials per participant than to increase the number of participants. While this might seem intuitive, as a higher number of games per participant translates into more data, this consideration is not traditionally included in power analyses. These results add to previous research on frequentist methods (Baker *et al.*, 2021; Rouder & Haaf, 2018), which suggest the inclusion of this parameter in design analyses.

It is important to highlight that while this analysis summarizes Bayes factors using common thresholds, their interpretation is not inherently discrete. The primary strength of Bayes factors lies in their continuous nature and in their role within a broader Bayesian framework that emphasizes principled model construction and the use of prior knowledge (Aczel *et al.*, 2020; Coventry & Bartlett, 2024; Gelman & Rubin, 1995; Gelman & Shalizi, 2013). However, in the present work, Bayes factors are used specifically as a pragmatic tool for design analysis, where threshold-based summaries aid

decision-making. In retrospective analyses, Bayes factors should still be interpreted within a full Bayesian framework rather than being reduced to a modified p-value.

Considerations about parameters such as sample size and the number of games per participant can be extended beyond the given example to studies measuring, for example, reaction times, EEG, MEG, or fMRI (Baker *et al.*, 2021; Lorenz *et al.*, 2017). If these analyses are considering latent variables, a suitable model of how this latent variable affects the outcome measure is needed. Previous studies in neuroscience have combined latent variable modelling with various methods including EEG (Ghaderi-Kangavari *et al.*, 2022; Li *et al.*, 2020), MRI (Cooper *et al.*, 2019; Lahey *et al.*, 2012; Tien *et al.*, 1996), reaction times (Jaffe *et al.*, 2023), and cognitive tasks (Decker *et al.*, 2014; Jaffe *et al.*, 2023). If a suitable model has not been established, a pilot study may provide one that can be used for the prospective design analysis.

The example used in this study is a simple version of an exploration task, using binary outcomes in a 2-choice paradigm. There are many different ways of targeting exploration, including using continuous outcomes (which is less straightforward for non-numerical outcomes such as pain), non-stationary paradigms (in which the outcome probabilities change over time), and larger numbers of options (e.g. 4 bandits). Another complexity is that humans use more than one type of exploration strategy (Seymour *et al.*, 2012), indeed the horizon task here was explicitly designed to explore so-called 'directed' exploration, which is proposed to operate over-and-above random exploration, according to estimates of outcome uncertainty (Wilson *et al.*, 2014, 2021). The approach we show here can be equally applied to these more complex paradigms and analyses (i.e. computational models) with consideration of runtime and scalability. In general, simulation ranges should focus on values that are theoretically meaningful, rather than the broader ranges used here for demonstration purposes. For highly complex designs, approximate methods may be necessary, and efficient implementation (e.g. vectorisation, parallel computing, or cluster use) becomes increasingly important.

In terms of potential limitations, the model used could be improved by modelling changes in exploration throughout a game or by extending the in-population variability to the learning rate. For the BFDA, rather than recording the total number of aversive outcomes for each game, an alternative approach would be to record the choice made in each trial and compute a likelihood function for choosing the respective arm for all possible Q-values for each trial. However, the resulting probabilities in the present approach regarding possible errors are still valuable, as they estimate appropriate sample sizes and number of games per participant. The dependence on an accurate generative model is a general limitation of using BFDA, as design recommendations are conditional on that model. If the true data-generating process deviates from the specified model, conclusions about the optimal design may be inaccurate. In settings with multiple competing computational models aiming to describe a process, Bayesian Design Optimisation can be used to compare candidate data-generating models (Melinscak and Bach, 2020). Independently, the adequacy of the analysis model can be evaluated empirically using experiment-based calibration (Bach & Melinscak, 2020; Bach *et al.*, 2020, 2023). However, if the goal is to determine which model in a candidate set best explains the data, BFDA can be extended to compare models directly. The Bayes factor quantifies the relative evidence between two models (M_i, M_j).

$$BF = \frac{p(Y|M_i)}{p(Y|M_j)}.$$

One approach would be to simulate data under each candidate model and examine, across sample sizes and other design parameters, when the Bayes factor reliably favours the true generative model. However, this procedure becomes increasingly complex as the number of competing models grows.

We argue that design analyses could benefit from calculating not only the optimal sample size and number of games per participant, but also from extending the analysis to other parameters. BFDA enables us to examine the effect of parameters such as the prior distribution or, more specific to the simulated task, the probabilities of aversive outcomes, the number of free choice trials, or the difference between continuous and binary outcomes. These considerations can help us make the best use of available resources.

Data availability

Underlying data

Zenodo: sarah407/BFDA_Multi-Armed-Bandit: v1.1, https://doi.org/10.5281/zenodo.18879706 (Schreiber, 2026).

This project contains the simulated data.

Code availability

Source code available from: https://github.com/sarah407/BFDA_Multi-Armed-Bandit

Archived source code at time of publication: https://doi.org/10.5281/zenodo.18879706 (Schreiber, 2026).

Acknowledgement

We used ChatGPT Edu (OpenAI; GPT-5.3) to support language editing during manuscript preparation.

References

Aczel B, Hoekstra R, Gelman A, *et al.*: **Discussion points for Bayesian inference.** *Nature Human Behaviour.* 2020; **4**(6): 561–563.
Publisher Full Text

Ashby D: **Bayesian statistics in medicine: a 25 year review.** *Stat Med.* 2006; **25**(21): 3589–3631.
PubMed Abstract | Publisher Full Text

Bach DR, Melinscak F: **Psychophysiological modelling and the measurement of fear conditioning.** *Behaviour Research and Therapy.* 2020; **127**: 103576.
PubMed Abstract | Publisher Full Text | Free Full Text

Bach DR, Melinščak F, Fleming SM, *et al.*: **Calibrating the experimental measurement of psychological attributes.** *Nature Human Behaviour.* 2020; **4**(12): 1229–1235.
PubMed Abstract | Publisher Full Text

Bach DR, Sporrer J, Abend R: **Consensus design of a calibration experiment for human fear conditioning.** *Neurosci Biobehav Rev.* 2023; **148**: 105146.
Publisher Full Text

Baker DH, Vilidaite G, Lygo FA, *et al.*: **Power contours: optimising sample size and precision in experimental psychology and human neuroscience.** *Psychol Methods.* 2021; **26**(3): 295–314.
PubMed Abstract | Publisher Full Text

Bayes T: **An essay towards solving a problem in the doctrine of chances. 1763.** *MD Comput.* 1991; **8**(3): 157–171.
PubMed Abstract

Benjamin DJ, Berger JO, Johannesson M, *et al.*: **Redefine statistical significance.** *Nat Hum Behav.* 2018; **2**(1): 6–10.
PubMed Abstract | Publisher Full Text

Blackwell D, Ramamoorthi RV: **A Bayes but Not Classically Sufficient Statistic.** *Ann Stat.* 1982; **10**(3): 1025–1026.
Publisher Full Text

Cao J, Lee JJ, Alber S: **Comparison of Bayesian sample size criteria: ACC, ALC, and WOC.** *J Stat Plan Inference.* 2009; **139**(12): 4111–4122.
PubMed Abstract | Publisher Full Text | Free Full Text

Cooper SR, Jackson JJ, Barch DM, *et al.*: **Neuroimaging of individual differences: a latent variable modeling perspective.** *Neurosci Biobehav Rev.* 2019; **98**: 29–46.
PubMed Abstract | Publisher Full Text

Coventry BS, Bartlett EL: **Practical Bayesian Inference in Neuroscience: Or How I Learned to Stop Worrying and Embrace the Distribution.** *eNeuro.* 2024; **11**(7): ENEURO.0484-0423.2024.
Publisher Full Text

Cronin B, Stevenson IH, Sur M, *et al.*: **Hierarchical Bayesian Modeling and Markov Chain Monte Carlo Sampling for Tuning-Curve Analysis.** *Journal of Neurophysiology.* 2010; **103**(1): 591–602.
PubMed Abstract | Publisher Full Text | Free Full Text

Czitrom V: **14. Introduction to Design of Experiments**. In *Statistical Case Studies for Industrial Process Improvement.* 2012; 171–198.
Publisher Full Text

Danwitz L, Mathar D, Smith E, *et al.*: **Parameter and model recovery of reinforcement learning models for restless bandit problems.** *Comput Brain Behav.* 2022; **5**(4): 547–563.
Publisher Full Text

Daw ND, O'Doherty JP, Dayan P, *et al.*: **Cortical substrates for exploratory decisions in humans.** *Nature.* 2006; **441**(7095): 876–879.
Publisher Full Text

Decker SL, Englund JA, Roberts AM: **Higher-order factor structures for the WISC-IV: implications for neuropsychological test interpretation.** *Applied Neuropsychology: Child.* 2014; **3**(2): 135–144.
PubMed Abstract | Publisher Full Text

Garud SS, Karimi IA, Kraft M: **Design of computer experiments: A review.** *Computers & Chemical Engineering.* 2017; **106**: 71–95.
Publisher Full Text

Gelman A: **Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper).** *Bayes Anal.* 2006; **1**(3): 515–534.
Publisher Full Text

Gelman A, Carlin J: **Beyond power calculations: assessing Type S (Sign) and Type M (Magnitude) errors.** *Perspect Psychol Sci.* 2014; **9**(6): 641–651.
PubMed Abstract | Publisher Full Text

Gelman A, Hill J: *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press; 2006.
Publisher Full Text

Gelman A, Rubin DB: **Avoiding Model Selection in Bayesian Social Research.** *Sociol Methodol.* 1995; **25**: 165–173.
Publisher Full Text

Gelman A, Shalizi CR: **Philosophy and the practice of Bayesian statistics.** *British Journal of Mathematical and Statistical Psychology.* 2013; **66**(1): 8–38.
PubMed Abstract | Publisher Full Text | Free Full Text

Gershman SJ: **Uncertainty and exploration.** *Decision.* 2019; **6**(3): 277–286.
PubMed Abstract | Publisher Full Text | Free Full Text

Ghaderi-Kangavari A, Rad JA, Parand K, *et al.*: **Neuro-cognitive models of single-trial EEG measures describe latent effects of spatial attention during perceptual decision making.** *J Math Psychol.* 2022; **111**: 102725.
Publisher Full Text

Gigerenzer G: **The irrationality paradox.** *Behav Brain Sci.* 2004; **27**(3): 336–338.
Publisher Full Text

Hudson TE: *Bayesian Data Analysis for the Behavioral and Neural Sciences: Non-Calculus Fundamentals.* Cambridge University Press; 2021.
Publisher Full Text

Ioannidis JPA, Greenland S, Hlatky MA, *et al.*: **Increasing value and reducing waste in research design, conduct, and analysis.** *Lancet.* 2014; **383**(9912): 166–175.
PubMed Abstract | Publisher Full Text

Jaffe PI, Poldrack RA, Schafer RJ, *et al.*: **Modelling human behaviour in cognitive tasks with latent dynamical systems.** *Nat Hum Behav.* 2023; **7**(6): 986–1000.
Publisher Full Text

Jeffreys H: **Some tests of significance, treated by the theory of probability.** *Math Proc Camb Philos Soc.* 1935; **31**(2): 203–222.
Publisher Full Text

Jeffreys H: *The theory of probability.* Oxford University Press; 1961.

Johnson VE, Pramanik S, Shudde R: **Bayes factor functions for reporting outcomes of hypothesis tests.** *Proc Natl Acad Sci U S A.* 2023; **120**(8): e2217331120.
PubMed Abstract | Publisher Full Text | Free Full Text

Joseph L, Bélisle P: **Bayesian sample size determination for normal means and differences between normal means.** *Journal of the Royal Statistical Society: Series D (The Statistician).* 1997; **46**(2): 209–226.
Publisher Full Text

Kass RE, Raftery AE: **Bayes factors.** *J Am Stat Assoc.* 1995; **90**(430): 773–795.
Publisher Full Text

Kruschke JK: **What to believe: Bayesian methods for data analysis.** *Trends in Cognitive Sciences.* 2010; **14**(7): 293–300.
Publisher Full Text

Kruschke JK: **Bayesian estimation supersedes the t test.** *J Exp Psychol Gen.* 2013; **142**(2): 573–603.
PubMed Abstract | Publisher Full Text

Kruschke JK: *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, second edition.* 2014.
Publisher Full Text

Kruschke JK: **Chapter 5 - Bayes' rule**. In: *Doing Bayesian Data Analysis.* (2 ed). Academic Press, 2015a; 99–120.
Publisher Full Text

Kruschke JK: **Chapter 10 - Model comparison and hierarchical modeling**. In: Kruschke JK: *Doing Bayesian Data Analysis.* (2 ed). Academic Press, 2015b; 265–296.
Publisher Full Text

Kruschke JK, Aguinis H, Joo H: **The time has come: Bayesian methods for data analysis in the organizational sciences.** *Organ Res Methods.* 2012; **15**(4): 722–752.
Publisher Full Text

Krypotos AM, Alves M, Crombez G, *et al.*: **The role of intolerance of uncertainty when solving the exploration-exploitation dilemma.** *Int J Psychophysiol.* 2022a; **181**: 33–39.
Publisher Full Text

Krypotos AM, Crombez G, Alves M, *et al.*: **The exploration-exploitation dilemma in pain: an experimental investigation.** *Pain.* 2022b; **163**(2): e215–e233.
Publisher Full Text

Krypotos AM, Crombez G, Vlaeyen JWS: **The dynamics of pain avoidance: the exploration-exploitation dilemma.** *Pain.* 2024; **165**: 2145–2149.
Publisher Full Text

Lahey BB, McNealy K, Knodt A, *et al.*: **Using confirmatory factor analysis to measure contemporaneous activation of defined neuronal networks in functional magnetic resonance imaging.** *Neuroimage.* 2012; **60**(4): 1982–1991.
PubMed Abstract | **Publisher Full Text**

Li X, Zhao Z, Song D, *et al.*: **Latent factor decoding of multi-channel EEG for emotion recognition through autoencoder-like neural networks.** *Front Neurosci.* 2020; **14**: 87.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Lorenz R, Hampshire A, Leech R: **Neuroadaptive Bayesian optimization and hypothesis testing.** *Trends Cogn Sci.* 2017; **21**(3): 155–167.
PubMed Abstract | **Publisher Full Text**

Macleod MR, Michie S, Roberts I, *et al.*: **Biomedical research: increasing value, reducing waste.** *Lancet.* 2014; **383**(9912): 101–104.
PubMed Abstract | **Publisher Full Text**

Melinscak F, Bach DR: **Computational optimization of associative learning experiments.** *PLOS Computational Biology.* 2020; **16**(1): e1007593.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Mizell JM, Wang S, Frisvold A, *et al.*: **Differential impacts of healthy cognitive aging on directed and random exploration.** *Psychol Aging.* 2024; **39**(1): 88–101.
PubMed Abstract | **Publisher Full Text**

Morey RD, Rouder JN: **Bayes factor approaches for testing interval null hypotheses.** *Psychol Methods.* 2011; **16**(4): 406–419.
PubMed Abstract | **Publisher Full Text**

Morrison DE, Henkel RE: *The significance test controversy: a reader*. Taylor and Francis; 2017.
Publisher Full Text

Munafò MR, Nosek BA, Bishop DVM, *et al.*: **A manifesto for reproducible science.** *Nature Human Behaviour.* 2017; **1**(1): 0021.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Pfister R: **Variability of Bayes factor estimates in bayesian Analysis of Variance.** *Quant Method Psychol.* 2021; **17**(1): 40–45.
Publisher Full Text

Rouder JN, Haaf JM: **Power, dominance, and constraint: a note on the appeal of different design traditions.** *Adv Methods Pract Psychol Sci.* 2018; **1**(1): 19–26.
Publisher Full Text

Ryan EG, Drovandi CC, McGree JM, *et al.*: **A review of modern computational algorithms for Bayesian optimal design.** *Int Stat Rev.* 2016; **84**(1): 128–154.
Publisher Full Text

Schönbrodt FD, Wagenmakers EJ: **Bayes Factor Design Analysis: planning for compelling evidence.** *Psychon Bull Rev.* 2018; **25**(1): 128–142.
PubMed Abstract | **Publisher Full Text**

Schreiber S: **sarah407/BFDA_Multi-Armed-Bandit: v1.1 (v1.1).** *Zenodo.* 2026.
Publisher Full Text

Seabold S, Perktold J: **Statsmodels: Econometric and Statistical Modeling with Python.** *SciPy 2010.* 2010.
Publisher Full Text

Seymour B, Daw ND, Roiser JP, *et al.*: **Serotonin selectively modulates reward value in human decision-making.** *J Neurosci.* 2012; **32**(17): 5833–5842.
PubMed Abstract | **Publisher Full Text**

Somerville LH, Sasse SF, Garrad MC, *et al.*: **Charting the expansion of strategic exploratory behavior during adolescence.** *J Exp Psychol Gen.* 2017; **146**(2): 155–164.
PubMed Abstract | **Publisher Full Text**

Stefan AM, Gronau QF, Schönbrodt FD, *et al.*: **A tutorial on Bayes Factor Design Analysis using an informed prior.** *Behav Res Methods.* 2019; **51**(3): 1042–1058.
PubMed Abstract | **Publisher Full Text**

Stefan AM, Gronau QF, Wagenmakers EJ: **Interim design analysis using Bayes factor forecasts.** *Psychol Methods.* 2024: 38330340.
Publisher Full Text

Storz-Pfennig P: **Potentially unnecessary and wasteful clinical trial research detected in cumulative meta-epidemiological and trial sequential analysis.** *J Clin Epidemiol.* 2017; **82**: 61–70.
PubMed Abstract | **Publisher Full Text**

Tien AY, Eaton WW, Schlaepfer TE, *et al.*: **Exploratory factor analysis of MRI brain structure measures in schizophrenia.** *Schizophr Res.* 1996; **19**(2–3): 93–101.
PubMed Abstract | **Publisher Full Text**

Tůmová O, Kupka L, Netolický P: **Design of Experiments approach and its application in the evaluation of experiments.** *2018 International Conference on Diagnostics in Electrical Engineering (Diagnostika).* 2018.

Valentin S, Kleinegesse S, Bramley NR, *et al.*: **Designing optimal behavioral experiments using machine learning.** *eLife.* 2024; **13**: e86224.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

van de Schoot R, Winter SD, Ryan O, *et al.*: **A systematic review of Bayesian articles in psychology: the last 25 years.** *Psychol Methods.* 2017; **22**(2): 217, 28594224–239.
Publisher Full Text

Van Dongen S: **Prior specification in Bayesian statistics: Three cautionary tales.** *Journal of Theoretical Biology.* 2006; **242**(1): 90–100.
PubMed Abstract | **Publisher Full Text**

Vize CE, Phillips NL, Miller JD, *et al.*: **On the Use and Misuses of Preregistration: A Reply to Klonsky (2024).** *Assessment.* 2025; **32**(2): 235–243.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Vlaeyen JWS, Crombez G, Linton SJ: **The fear-avoidance model of pain.** *Pain.* 2016; **157**(8): 1588–1589.
PubMed Abstract | **Publisher Full Text**

Wagenmakers EJ: **A practical solution to the pervasive problems of p values.** *Psychon Bull Rev.* 2007; **14**(5): 779–804.
PubMed Abstract | **Publisher Full Text**

Wagenmakers EJ, Marsman M, Jamil T, *et al.*: **Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications.** *Psychon Bull Rev.* 2018; **25**(1): 35–57.
PubMed Abstract | **Publisher Full Text**

Waltz JA, Wilson RC, Albrecht MA, *et al.*: **Differential effects of psychotic illness on directed and random exploration.** *Comput Psychiatr.* 2020; **4**: 18–39.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wang O, Lee SW, O'Doherty J, *et al.*: **Model-based and model-free pain avoidance learning.** *Brain and Neuroscience Advances.* 2018; **2**: 2398212818772964.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Wilson RC, Bonawitz E, Costa VD, *et al.*: **Balancing exploration and exploitation with information and randomization.** *Curr Opin Behav Sci.* 2021; **38**: 49–56.
PubMed Abstract | **Publisher Full Text**

Wilson RC, Geana A, White JM, *et al.*: **Humans use directed and random exploration to solve the explore-exploit dilemma.** *J Exp Psychol Gen.* 2014; **143**(6): 2074–2081.
PubMed Abstract | **Publisher Full Text**

Zhang S, Lee MD: **Optimal experimental design for a class of bandit problems.** *J Math Psychol.* 2010; **54**(6): 499–508.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ? ? ✓ ?

Version 2

Reviewer Report 05 June 2026

https://doi.org/10.21956/wellcomeopenres.28844.r155003

✓ **Jessica Schaaf**
[1] Medical Neuroscience, adboud University Donders Institute for Brain Cognition and Behaviour (Ringgold ID: 198328), Nijmegen, Gelderland, The Netherlands
[2] Radboudumc (Ringgold ID: 6034), Nijmegen, Gelderland, The Netherlands

I thank the authors for their revisions acknowledging the boundaries of their method and referring to additional material where necessary. I do have some remaining comments.

In response to point 1, you have extended the README file in the Github repository. Although it already helps to know what every script does and what it outputs, the README file still doesn't state *why* I would run a given script and based on which output I should conclude which design to use. I think it would be a shame if researchers do not apply your elegant method because it requires too much cognitive effort.

In response to point 7, you have added the coding of outcomes. I understand the combination of the coding of outcomes and the initialization of Q-values doesn't matter in simulation studies as long as you simulate and refit with the same coding. However, in empirical data, it is common to initialize the Q-value in between the possible outcomes (in your case -0.5). Otherwise you assume participants are biased towards expecting no aversive outcome a priori, which affects learning rate estimation. It would be good to explicitly state the rationale for choosing this Q-value initialization (or simply state it was an arbitrary choice).

In response to point 8, you adjusted Figures 3 and 4. Really helpful. I don't see green circles in Figure 3c and d though.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian computational modeling; reinforcement learning; decision making; cognitive development

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 05 June 2026

https://doi.org/10.21956/wellcomeopenres.28844.r155582

?

Tommaso Costa
University of Turin, Turin, Italy

This manuscript presents a tutorial-style methodological framework combining Bayes factor design analysis (BFDA) with latent variable modelling in the context of multi-armed bandit (MAB) tasks. The article addresses an important and timely problem in cognitive neuroscience and behavioural science, namely how to prospectively optimise experimental designs when the quantities of interest are latent computational parameters rather than directly observed variables.

The manuscript is clearly motivated and generally well written. A major strength of the work is the integration of BFDA with computational modelling approaches commonly used in reinforcement learning and decision neuroscience. The paper also contributes by explicitly considering probabilities of inferential errors, uncertainty estimates, and practical trade-offs between sample size and number of trials per participant. The open availability of code and data further strengthens the transparency and reproducibility of the work.

Overall, I believe the article makes a useful methodological contribution and is suitable for indexing after revision. My main concerns relate not to the general validity of the framework, but rather to the interpretation, generalisability, and robustness of the presented implementation.

Major comments

1. Parameter identifiability and simplicity of the computational model

One important limitation is that the current demonstrations rely on a relatively simplified computational model. The analyses focus primarily on a single softmax exploration parameter τ with a fixed learning rate. While this simplification is understandable for tutorial purposes, it substantially reduces the complexity typically encountered in realistic reinforcement-learning analyses.
In many practical bandit paradigms, exploration behaviour depends on multiple partially correlated latent processes, including directed exploration, uncertainty bonuses, perseveration, or variable learning rates. Under these conditions, parameter identifiability can become substantially more difficult.
The manuscript demonstrates encouraging recovery properties using MSEE analyses and regression summaries, but additional discussion of parameter identifiability would strengthen the methodological contribution. In particular, the authors should discuss more explicitly:

- parameter trade-offs,
- posterior parameter correlations,

- and the extent to which the present conclusions are expected to generalise to richer computational models.

Although full additional analyses may not be strictly necessary for publication, a more explicit discussion of these limitations is important to ensure appropriate interpretation of the framework.

1. Prior sensitivity analysis

The manuscript correctly acknowledges the importance of prior specification in Bayesian inference. However, given that Bayes factors can be highly sensitive to prior assumptions, it is somewhat surprising that the paper does not include a more systematic prior sensitivity analysis. The current demonstrations rely primarily on uniform priors for illustrative purposes. While this is acceptable for a tutorial paper, the manuscript would benefit from either:

- a brief sensitivity analysis across different plausible priors,
  or
- a more detailed discussion of how prior specification may affect BFDA conclusions in practical applications.

This point is particularly important because BFDA is intended as a prospective design tool, and design recommendations may depend substantially on prior assumptions.

1. Interpretation of Bayes factor thresholds

The manuscript appropriately notes that Bayes factors are continuous measures of evidence and should not be interpreted simply as Bayesian analogues of p-values. However, much of the analysis is nevertheless organised around threshold-based interpretations such as BF > 10 or BF < 1/3.

While this approach is understandable for pragmatic design-analysis purposes, the manuscript would benefit from a slightly deeper discussion clarifying:

- the pragmatic role of these thresholds,
- their relationship to decision-making,
- and the limitations of threshold-based evidence categorisation.

At present, the framework risks partially reproducing a threshold-centred logic analogous to traditional significance testing. Clarifying this distinction would strengthen the conceptual framing of the paper.

1. Generalisability and scalability

The manuscript demonstrates the framework using a relatively simple binary two-armed bandit task. Although the authors discuss possible extensions to more complex paradigms, the practical scalability of the method remains somewhat unclear.

It would be useful for the manuscript to elaborate further on:

- computational burden,
- scalability to hierarchical models with multiple latent parameters,
- runtime considerations,
- and applicability to more complex task structures.

This additional discussion would improve the translational relevance of the framework for researchers working with more realistic behavioural and neurocomputational models.

Minor comments

1. Some notation could be clarified further, particularly regarding the distinction between participant-level latent parameters and population-level parameters.
2. A brief explanation of why a fixed-n BFDA framework was prioritised over sequential BFDA approaches would be useful.

3. A final language-polishing pass would improve readability in some sections.
4. Some figures are visually dense and may benefit from slightly larger labels or simplified visual presentation.

Overall assessment

This is a thoughtful and methodologically valuable contribution that extends BFDA into an important area of computational cognitive neuroscience. The manuscript is transparent, technically competent, and addresses practically relevant design problems that are often neglected in behavioural research.

The integration of latent-variable modelling with prospective Bayesian design analysis is particularly valuable, and the emphasis on uncertainty, inferential errors, and design trade-offs is a major strength of the paper.

I recommend approval with reservations. The points above should be addressed to strengthen the robustness,interpretability, and generalisability of the proposed framework.

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian Statistics; Research Methodology; Cognitive Neuroscience

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 16 December 2024

https://doi.org/10.21956/wellcomeopenres.24565.r112332

?

Jessica Schaaf
[1] Medical Neuroscience, adboud University Donders Institute for Brain Cognition and Behaviour (Ringgold ID: 198328), Nijmegen, Gelderland, The Netherlands
[2] Radboudumc (Ringgold ID: 6034), Nijmegen, Gelderland, The Netherlands

The authors provide a Bayesian approach to do prospective planning of experimental studies including computational modeling. They illustrate their approach for a horizon task and do extensive simulations showing how to choose a sample size and the number of games per participant for different scenarios. I personally enjoyed reading the paper and think it provides a clear story on how to use Bayesian methods to optimize experimental research.

Below I provide some suggestions on how to improve the applicability and clarity of the method. As these are mostly clarifications and extensions, I consider them minor revisions.

1. My most important point is the correspondence between the target audience and the information provided. I assume the paper is meant for the applied researcher wishing to use the proposed method to plan their experiment. I applaud you from sharing your code, but I would not know where to start. Which considerations do I have to make *before* performing the simulations? What kind of range would I have to simulate? Which script(s) do I need to adjust? What are the software requirements for using the code? It would really help to have a step-by-step starters guide, either by upping the tutorial element of your paper or by providing a guide accompanying your code base.

2. One advantage of the two-armed experiment with fixed horizon is that it is fairly easy to map all possible choice and outcome patterns, and thus their likelihoods. Although you state that “the approach we show here can be equally applied to these more complex paradigms and analyses (i.e. computational models)”, I struggle to imagine how feasible the approach is in such cases as the number of possible patterns increases exponentially with more choice options and more trials. Please elaborate.

3. Somewhat relatedly, in the discussion you state that “the dependence on an accurate model is a drawback of using BFDA with latent variables and restricts its uses as a prospective design analysis in cases where a model is missing.” Although this is true, I wonder how this affects the applicability of your method. Say I want to test which model in a model set fits my data best (as is recommended practice, see e.g., Wilson & Collins, 2019; Van den Bos et al., 2018). Would this mean that I have to do a BFDA for all my competing models? And how do I then decide which sample size and number of games per participant to use?

4. Bayes Factor rely heavily on the chosen prior. Although you mention this yourselves in the method (“The chosen prior can heavily influence the outcome of an analysis and should

therefore be chosen carefully, as it will bias estimation.") and discussion ("In the example we used an uninformative prior, but depending on the experiment and existing literature, this should be adjusted. A well-informed prior can lead to higher probabilities of detecting a difference between groups."), I miss explanation on how it exactly does so. It would be even better to also explicate considerations with respect to prospective planning.

5. In general, I find the paper clearly written. Yet, I got lost in the method section. Using a numerical example alongside the formulas would make the method more graspable.

6. Thank you for mentioning how you initialized Q-values (it is important!). Please also add how you coded rewards enabling readers to assess the accuracy of the initialization.

7. I really like the numerical examples (i.e., the combination of sample size, number of games, and probability of obtaining a certain BF) you used to illustrate the results and the beautiful plots. Yet, I struggled to connect the two. To guide readers, I think it would help to, for example, mark in the plot where the values in the text came from.

References

1. Robert C Wilson, Anne Ge Collins: Ten simple rules for the computational modeling of behavioral data. Publisher Full Text
2. Wouter van den Bos, Rasmus Bruckner, Matthew R Nassar, Rui Mata, Ben Eppinger,: Computational neuroscience across the lifespan: Promises and pitfalls. Publisher Full Text

Is the rationale for developing the new method (or application) clearly explained?

Yes

Is the description of the method technically sound?

Yes

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bayesian computational modeling; reinforcement learning; decision making; cognitive development

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Apr 2026

Sarah Schreiber

The authors provide a Bayesian approach to do prospective planning of experimental studies including computational modeling. They illustrate their approach for a horizon task and do extensive simulations showing how to choose a sample size and the number of games per participant for different scenarios. I personally enjoyed reading the paper and think it provides a clear story on how to use Bayesian methods to optimize experimental research.
Response: We are thankful to Reviewer 3 for a generally positive evaluation of our work.

Point 1 & 2.
My most important point is the correspondence between the target audience and the information provided. I assume the paper is meant for the applied researcher wishing to use the proposed method to plan their experiment. I applaud you from sharing your code, but I would not know where to start. Which considerations do I have to make before performing the simulations? What kind of range would I have to simulate? Which script(s) do I need to adjust? What are the software requirements for using the code? It would really help to have a step-by-step starters guide, either by upping the tutorial element of your paper or by providing a guide accompanying your code base.
Response: We thank the reviewer for this helpful comment. We agree that a step-by-step guide would make the manuscript more accessible. However, including a detailed tutorial within the manuscript itself would substantially increase its length and risk overwhelming readers with implementation details. Therefore, to address this concern, we have expanded the documentation in the GitHub repository (https://github.com/sarah407/BFDA_Multi-Armed-Bandit). Specifically, we have added clearer descriptions of the analysis scripts, outlined the software requirements, and included a conceptual checklist to guide users in setting up and running their own simulations. We have also provided references to additional resources for readers seeking more detailed guidance.

Point 3.
One advantage of the two-armed experiment with fixed horizon is that it is fairly easy to map all possible choice and outcome patterns, and thus their likelihoods. Although you state that "the approach we show here can be equally applied to these more complex paradigms and analyses (i.e. computational models)", I struggle to imagine how feasible the approach is in such cases as the number of possible patterns increases exponentially with more choice options and more trials. Please elaborate.
Response: We thank the reviewer for this important point. We agree that as task complexity increases (e.g., more options, longer horizons, or latent-variable models), the number of possible choice-outcome patterns grows rapidly, making exhaustive likelihood mapping computationally demanding. Our intention was to highlight that the general simulation-based framework can, in principle, be extended to more complex paradigms, such as MAB with more arms, dynamically changing reward probabilities, or hierarchical modelling, but

only with careful consideration of computational constraints. In such cases, researchers would typically restrict the parameter space to theoretically relevant ranges, focus on key parameters, and limit the sample size grid. We have clarified in the manuscript that applying the approach to more complex designs requires explicit attention to runtime and scalability.
Revised Section: "The approach we show here can be equally applied to these more complex paradigms and analyses (i.e. computational models) with consideration of runtime and scalability. In general, simulation ranges should focus on values that are theoretically meaningful, rather than the broader ranges used here for demonstration purposes. For highly complex designs, approximate methods may be necessary, and efficient implementation (e.g. vectorisation, parallel computing, or cluster use) becomes increasingly important.**"**

Point 4.
Somewhat relatedly, in the discussion you state that "the dependence on an accurate model is a drawback of using BFDA with latent variables and restricts its uses as a prospective design analysis in cases where a model is missing." Although this is true, I wonder how this affects the applicability of your method. Say I want to test which model in a model set fits my data best (as is recommended practice, see e.g., Wilson & Collins, 2019; Van den Bos et al., 2018). Would this mean that I have to do a BFDA for all my competing models? And how do I then decide which sample size and number of games per participant to use?
Response: We agree that this increases the complexity of the approach rather than fundamentally limiting it. Model comparison remains possible, but it requires additional simulation steps and clearly defined decision criteria. We have elaborated on this scenario in the revised manuscript to clarify how it can be implemented in practice.
Revised Section: "However, if the goal is to determine which model in a candidate set best explains the data, BFDA can be extended to compare models directly. The Bayes factor quantifies the relative evidence between two models (M_i, M_j) $BF = \frac{p(Y|M_i)}{p(Y|M_j)}$. One approach would be to simulate data under each candidate model and examine, across sample sizes and other design parameters, when the Bayes factor reliably favours the true generative model. However, this procedure becomes increasingly complex as the number of competing models grows."

Point 5.
Bayes Factor rely heavily on the chosen prior. Although you mention this yourselves in the method ("The chosen prior can heavily influence the outcome of an analysis and should therefore be chosen carefully, as it will bias estimation.") and discussion ("In the example we used an uninformative prior, but depending on the experiment and existing literature, this should be adjusted. A well-informed prior can lead to higher probabilities of detecting a difference between groups."), I miss explanation on how it exactly does so. It would be even better to also explicate considerations with respect to prospective planning.
Response: While an extensive treatment and comparison of prior specification is beyond the scope of this tutorial, we have added references to relevant literature to provide readers with further guidance.
Revised Section: "The chosen prior can heavily influence the outcome of an analysis and

should therefore be chosen carefully, as it will bias the estimation. When prior knowledge about the underlying distribution is limited, researchers often employ uninformative or weakly informative priors, which can nevertheless reflect plausible bonds for the parameters. Given the subjective nature of priors, their selection should be transparently reported and sufficiently motivated. In addition, sensitivity analyses can be conducted to evaluate the robustness of results with respect to the chosen prior. For further discussions on priors, refer to further literature such as Gelman (2006), Van Dongen (2006), or Stefan (2019)."

Point 6.

In general, I find the paper clearly written. Yet, I got lost in the method section. Using a numerical example alongside the formulas would make the method more graspable.

Response: We thank the reviewer for this suggestion. We agree that numerical examples can aid understanding, particularly in tutorial-style presentations. However, providing worked numerical examples for all formal expressions would require substantial restructuring of the Methods section and would considerably increase its length. We have revised the Methods section to improve clarity and readability, including clearer transitions and additional explanatory text where appropriate. We also provide references to established textbooks and methodological resources for readers seeking more step-by-step or example-based guidance.

Added Section: ***"Bayesian statistics.*** We begin by outlining the central concepts and key formulas of Bayesian statistics. Readers seeking more detailed explanations, formal derivations, and worked examples are referred to Kruschke (2014) and Hudson (2021)."

Point 7.

Thank you for mentioning how you initialized Q-values (it is important!). Please also add how you coded rewards enabling readers to assess the accuracy of the initialization.

Response: We thank the reviewer for this comment and have adjusted the text accordingly.

Revised Section: "In the computational modelling of these outcomes, each outcome must be assigned a numerical value. Here, the outcome of each choice was coded as 0 or -1, where -1 represents an aversive outcome.**"**

Point 8.

I really like the numerical examples (i.e., the combination of sample size, number of games, and probability of obtaining a certain BF) you used to illustrate the results and the beautiful plots. Yet, I struggled to connect the two. To guide readers, I think it would help to, for example, mark in the plot where the values in the text came from.

Response: We thank the reviewer for this helpful suggestion. To improve clarity and better connect the numerical examples in the text with the corresponding figures, we have revised the figures 3 and 4 to explicitly mark the parameter combinations referenced in the manuscript (e.g., specific sample sizes and numbers of games per participant).

Revised Section: "Using a rough estimate of $\mu_\tau \approx 0.1$, we can infer from Figure 3a that about 40 participants per group would be needed to achieve an above 50% chance of the analysis yielding Bayes factors above 10, highlighted with a white circle." "The exact percentages reported here are obtained directly from our simulations, while the broader probability regions are illustrated in the contour plot (Figure 3)." "Alternatively, a sample size of 30, with each participant completing 130 games could be implemented (Figure 3c, green circles), resulting in a smaller total amount of games played. For both options the probability

of wrongfully supporting the null hypothesis is below 2% (Figure 3d, green circles).” “To illustrate further, we can consider our example from the previous section, where we compared n = 30 with n_{games} = 130 to n = 20 with n_{games} = 200. The probability of obtaining a Bayes factor greater than 10 was 90% for both options. To decide between the two options, we can consider the magnitude error. The first option has close to a 35% probability of overestimating the effect by at least 10% for all significant results. For the latter, this probability is between 30% and 35% (Figure 4a, green circles). Therefore, the latter option is less likely to overestimate the true effect size, although these probabilities are still relatively high.”

Competing Interests: No competing interests were disclosed.

Reviewer Report 26 November 2024

https://doi.org/10.21956/wellcomeopenres.24565.r95853

?

Dominik Bach
University of Bonn, Bonn, Germany

Schreiber et al. give a tutorial-style introduction into, and worked example of, the established framework of Bayes Factor Design Analysis, applied in the context of a multi-armed bandit task. No new method is developed. Thus, the original scientific contribution appears relatively small, but the worked example provides a welcome illustration of this (suite of) method(s) for students of the field. My major concern is that two assumptions of the method are not sufficiently discussed; there are also a couple of minor technical concerns.

1. Any results hinge on the assumption of a specific data-generating process. If the true data-generating process (in a future experiment) is not the same as the one assumed in the simulation, then results such as required group size etc. will be misleading. The authors might want to refer to (Melinscak and Bach, 2020) as an example of how to use Bayesian Design Optimisation to plan experiments to decide between data-generating models, rather than parameters of the same model in different participant groups, as done here.

2. The results also hinge on the (somewhat related but not equivalent) assumption that the data-generating process is exactly the same as the model used to analyse the data (i.e. the measurement model). The appropriateness of the measurement model can be assessed wihtout knowing the data-generating process, by running an experiment with a manipulation the effect of which is (somewhat) known, and testing different data analysis methods in terms of how well they reproduce the known effect. This method is known as experiment-based calibration (Bach et al., 2023, 2020), and the relevant metric as retrodictive validity (Bach and Melinscak, 2020).

3. The notation of many integrals is incorrect; sometimes the integration variable („dtheta“) is not stated (e.g. eq. 1-2), often the integration interval is missing or incorrectly replaced with the integration variable (e.g. eq. 1 should read int_{integration_interval} rather than int_{integration_variable}.).

4. It is rather confusing that the same symbol P is used to refer to multiple functions depending on what argument they take. That is, P(theta) is supposed to be something different from, let's say, P(omega). This is an abuse of mathematical notation. A common and useful convention in mathematical statistics is to differentiate probability distributions by subscript (e.g. P_X|Y, P_theta, P_omega, etc.). Psychological statistics literature is obviously rife with this abuse of notation, but in a sophisticated and heavily mathematical manuscript like this, the authors can clearly do better.

References

1. Dominik R. Bach, Filip Melinscak: Psychophysiological modelling and the measurement of fear conditioning. Publisher Full Text
2. Dominik R. Bach ,Juliana Sporrer,Rany Abend,Tom Beckers,Joseph E. Dunsmoor,Miquel A. Fullana,Matthias Gamer,Dylan G. Gee,Alfons Hamm,Catherine A. Hartley,Ryan J. Herringa,Tanja Jovanovic,Raffael Kalisch,David C. Knight,Shmuel Lissek,Tina B. Lonsdorf,Christian J. Merz, Mohammed Milad, Jayne Morriss, Elizabeth A. Phelps, Daniela Schiller: Consensus design of a calibration experiment for human fear conditioning. Publisher Full Text
3. Melinscak, F., Bach, D.R., 2020. Computational optimization of associative learning experiments. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1007593.

Is the rationale for developing the new method (or application) clearly explained?

No

Is the description of the method technically sound?

Partly

Are sufficient details provided to allow replication of the method development and its use by others?

Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Apr 2026

Sarah Schreiber

Schreiber et al. give a tutorial-style introduction into, and worked example of, the established framework of Bayes Factor Design Analysis, applied in the context of a multi-armed bandit task. No new method is developed. Thus, the original scientific contribution appears relatively small, but the worked example provides a welcome illustration of this (suite of) method(s) for students of the field. My major concern is that two assumptions of the method are not sufficiently discussed; there are also a couple of minor technical concerns.
Response: We are thankful to Reviewer 2 for a generally positive evaluation of our work.

Point 1 & 2:
Any results hinge on the assumption of a specific data-generating process. If the true data-generating process (in a future experiment) is not the same as the one assumed in the simulation, then results such as required group size etc. will be misleading. The authors might want to refer to (Melinscak and Bach, 2020) as an example of how to use Bayesian Design Optimisation to plan experiments to decide between data-generating models, rather than parameters of the same model in different participant groups, as done here.
The results also hinge on the (somewhat related but not equivalent) assumption that the data-generating process is exactly the same as the model used to analyse the data (i.e. the measurement model). The appropriateness of the measurement model can be assessed wihtout knowing the data-generating process, by running an experiment with a manipulation the effect of which is (somewhat) known, and testing different data analysis methods in terms of how well they reproduce the known effect. This method is known as experiment-based calibration (Bach et al., 2023, 2020), and the relevant metric as retrodictive validity (Bach and Melinscak, 2020).
Response: We thank the reviewer for highlighting this important point. In response, we have expanded the limitations section to more clearly discuss the dependence of BFDA on the assumed generative model. We now also explicitly refer to Bayesian Design Optimisation as an approach for comparing competing data-generating models, and to experiment-based calibration for evaluating the validity of the measurement model.
Revised Section: "The dependence on an accurate generative model is a general limitation of using BFDA, as design recommendations are conditional on that model. If the true data-generating process deviates from the specified model, conclusions about the optimal design may be inaccurate. In settings with multiple competing computational models aiming to describe a process, Bayesian Design Optimisation can be used to compare candidate data-generating models (Melinscak and Bach, 2020). Independently, the adequacy of the analysis model can be evaluated empirically using experiment-based calibration (Bach & Melinscak, 2020; Bach et al., 2020; Bach et al., 2023)."

Point 3.
The notation of many integrals is incorrect; sometimes the integration variable („dtheta") is not stated (e.g. eq. 1-2), often the integration interval is missing or incorrectly replaced with the integration variable (e.g. eq. 1 should read int_{integration_interval} rather than int_{integration_variable}.).
Response: We thank the reviewer for pointing this out and have revised the equations accordingly.

Point 4.
It is rather confusing that the same symbol P is used to refer to multiple functions depending on what argument they take. That is, P(theta) is supposed to be something different from, let's say, P(omega). This is an abuse of mathematical notation. A common and useful convention in mathematical statistics is to differentiate probability distributions by subscript (e.g. P_X|Y, P_theta, P_omega, etc.). Psychological statistics literature is obviously rife with this abuse of notation, but in a sophisticated and heavily mathematical manuscript like this, the authors can clearly do better.
Response: We thank the reviewer for this helpful comment. We have revised the notation throughout the manuscript to avoid overloading the symbol P and making clear when probability distributions are being used.
References 1. Dominik R. Bach, Filip Melinscak: Psychophysiological modelling and the measurement of fear conditioning. | Publisher Full Text 2. Dominik R. Bach ,Juliana Sporrer,Rany Abend,Tom Beckers,Joseph E. Dunsmoor,Miquel A. Fullana,Matthias Gamer,Dylan G. Gee,Alfons Hamm,Catherine A. Hartley,Ryan J. Herringa,Tanja Jovanovic,Raffael Kalisch,David C. Knight,Shmuel Lissek,Tina B. Lonsdorf,Christian J. Merz, Mohammed Milad, Jayne Morriss, Elizabeth A. Phelps, Daniela Schiller: Consensus design of a calibration experiment for human fear conditioning. | Publisher Full Text Melinscak, F., Bach, D.R., 2020. Computational optimization of associative learning experiments. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1007593.

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 October 2024

https://doi.org/10.21956/wellcomeopenres.24565.r104233

?

Brandon S Coventry
University of Wisconsin-Madison, Madison, Wisconsin, USA

Dear authors,

It was a pleasure to read your article detailing Bayes Factor design of experiments and its incorporation into latent variable modeling. Much well deserved interest has recently been received to incorporate Bayesian methods to the design of experiments, and I'm pleased to see your method adding to the conversation. I believe that modeling experiments during the experimental design phase with well-defined models of decision processes can certainly help bound and inform the BFDA process in meaningful ways. However, there are several areas in the manuscript that require further refinement to enhance both its readability and the broader adoption of your model and methodology.

I have outlined my feedback below, categorizing my suggestions into general, article-wide comments, as well as specific line-by-line revisions tied to particular sections of the text.

General Comments
I believe that the description of the MAB task needs much better description. As it stands, it's placed in the context of a pain model that is never described, and the definition of a painful stimulus is vague at best.

I also believe that this article often verges into vague descriptions of what is being done. I've tried to highlight sections below, but I believe portions can be rewritten to explicitly state what is being done. This will greatly increase readability.

It should be noted that the use of Bayes factors in model comparisons can be tricky, with the potential to reduce meaningful Bayesian inference to a modified p-value. Many argue that the power of Bayesian thinking lies in using prior knowledge to build the best possible models and to avoid the temptation to just compare every model available. For a good discussion of these caveats, see (Gelman and Rubin, 1995; Gelman and Shalizi, 2013). Providing a discussion of and context to BFDA may help allay some of these criticisms of BF analyses.

I also believe this article would benefit greatly from following the Bayesian analysis reporting guidelines (BARG) for article clarity and reproducibility (Kruschke, 2021).

Specific Comments

Abstract
Interest in the use of Bayesian statistics I believe extends well beyond just the fact that one can incorporate prior knowledge. While that is certainly a powerful component of Bayesian design, other advantageous features include the fact that inference is performed using the data on hand without implicit assumptions on population distributions, the ability to easily handle data distributions which are not strictly normal, direct quantification of uncertainty under inference, and the ability to iteratively update hypotheses as data is observed.

Introduction
"Designing an effective experiment can be challenging due to the number of parameters that need to be considered." While this is true, I believe this is a bit reductive, potentially to the point of trivializing design of experiments (DOE). Rewording of text and adding appropriate discussion of the goals of DOE would add important context to the reader, especially those relatively new to the science and admittedly art of DOE.

"To improve a study's effective-ness, it is recommended to conduct a thorough prospective design analysis rather than relying purely on a retrospective approach".
While true, I believe it's important to mention preregistration as part of the current initiatives to improve reproducibility. For a thorough discussion of the pros and cons of preregistration, see (Vize et al., 2024).

"Bayesian methods are gaining recognition for their advantages, as they allow the incorporation of prior knowledge into statistical processes."
Again, I think much more can be said about the advantages of Bayesian approaches. See

comments above in abstract section, and the following references (Blackwell and Ramamoorthi, 1982; Gelman and Shalizi, 2013; Kruschke, 2013, 2010).

“Traditionally, sample size determination methods rely on frequentist statistics, but there has been an ongoing critique among statisticians and methodologists regarding these approaches.”
While true, I’d argue that the primary goal of DOE is not just to determine sample size, but to determine sample size to ensure an experiment has the statistical power to detect if a desired effect is truly present. Sample size is a means to this end of the question experimenters are truly interested in.

“a-priori”
Please change to *a priori*

“BFDA assumes a population with predetermined attributes and conducts simulations on repeated sampling from this population”
In the context of a tutorial article, I believe concrete examples are going to be needed to best orient the reader to the goals of BFDA. One way to do this is to provide a quick hypothetical experiment, outlining what example predetermined attributes may be, and how one might conduct simulations. Are there experiments where simulations are generally widely available and accepted? What about cases of a novel experiment, how might one deal with model selection? These are questions that should be addressed to best help the reader adopt this framework.

“However, in psychology and neuroscience, we often deal with latent variables that first need to be inferred from the data collected.”
Absolutely, and I’m glad this is mentioned. It should be noted that in neuroscience and social sciences, initial passes where latent variables are truly hidden is often addressed with hierarchical Bayesian models. See (Cronin et al., 2010; Gelman, 2006; Gelman and Hill, 2006).

Methods
“The prior distribution P(θ) represents our knowledge about the distribution of θ ahead of data acquisition. The chosen prior can heavily influence the outcome of an analysis and should therefore be chosen carefully, as it will bias the estimation.”
True, especially if the observed data is weak, in which case priors dominate. However, some guidance of proper prior choice should be detailed in this article. Potentially adding a prior predictive checks to at least one model to help inform the reader of exactly this condition.

“Once the likelihoods have been calculated for each group, we can compare θ between two groups, for example between patients and healthy controls.”
Need to clarify for the reader whether you mean a point estimate of θ or the distribution of θ.

“Figure 1 panel A”
I think this figure has the promise of being very good, but is a bit confusing to read as it stands. As an example, the box “Compute the posterior” is positioned right next to, what I believe is the incorporation of the prior into the likelihood to create the posterior, but it appears based on where the box is positioned that you mean to say P(theta) is the posterior.

“HDPI is a method of obtaining a credible interval for skewed probability distributions.”
Need to describe credible regions, and I believe you need to redo the definition of the HPDI, as it’s

not just for skewed distributions. I think I understand what you mean by skewed here, but as it stands its ambiguous given that one can easily define HPDIs for non-skewed distributions.

“Figure 1”
Consider reorganizing the panels of Figure 1 as they are referenced out of order in text, making the figure a little harder to read.

“The probability of receiving a certain number of painful stimuli in one game is dependent on the choices of the participant on one hand and the reward probabilities of each arm on the other hand.”
So this is verging into confusion. I understand that MAB problems are used in models of pain as mentioned above, but if this terminology is going to be adapted in the method (painful stimuli), then this model needs to be fleshed out. Otherwise, I don’t not have a well defined view of what MAB model parameters, such as reward and painful stimuli mean. I’d suggest either removing the pain model from the MAB context and explicitly defining model rewards and failures or totally rope the description of the MAB problem into the pain model context.

“The learning rate α was fixed at 0.1”
This seems a bit high of a learning rate value. Why was this chosen and if there is appropriate literature, cite it here.

“Therefore, there are seven possible outcomes, as the agent can receive between zero and six painful stimuli.”
Again, this needs much better defining.

“For each choice, there are four possible combinations of arm chosen and aversive or neutral reward. For six free choice trials there are a total of 46 possible paths, or combinations of choices and outcomes.”
Moving this earlier in the text will help with understanding of the MAB model.

Results
“The use of the BFDA as a method for prospective design analysis allows us to consider a number of factors. However, it is essential to evaluate the algorithm estimating the latent variable as a preliminary step.”
This needs to be contextualized for understanding. What factors are you considering? What latent variables are you hoping to describe? What specific algorithm do you mean here? This sentence is quite vague. Being quite specific in what results you are showing will help readability greatly.

“Once the algorithm has been validated, it can be used for the BFDA, in which repeated sampling from a population is simulated and the planned analysis carried out on the sample.”
A descriptive flow chart figure would be extremely useful here.

“The individual exploration parameter for each simulated participant was drawn from a normal distribution with population mean μτ”.
Point to the exact parameter referenced in the methods, as right now this is vague.

“The overall fit of the regression model was statistically significant ($F(3,496) = 136300$, $p < 0.001$, $R2 = 0.999$)”

This regression model isn't described in the methods. Please detail what exactly is done here. It's also unclear why a frequentist approach here was used to validate the algorithm, given that the aim was a Bayes factor approach. This needs better detailing for why this was done.

"250 simulations were run for each combination of Δμτ ranging from 0 to 1 in increments of 0.05 and sample size per group ranging from 10 to 60 in increments of 2."
Please make this a little clearer. What simulation? Markov decision process? Is this MAB RL? Better descriptions of what is being done need to be added.

"Figure 2"
This is interestingly. What is happening at 0.05? It seems that BF dominate predictions except in this small window. When there's 0 difference in the mean, Bayes is doing create, but minimal and not moderate uncertainty proves problematic?

"For Δμτ ≠ 0 this probability is the probability of wrongfully accepting the null hypothesis, which corresponds to the type II error."
In Bayesian models, type I/II errors are generally not considered for a variety of reasons related to how Bayesian inference is performed. Bayesian approaches instead talk of Type M or Type S errors generally. You can derive a type I/II error for Bayesian approaches, but care must be taken in doing so. See (Gelman and Carlin, 2014).

"Figure 3"
Please label color map axis in this figure.

"This is supported by positive correlations between P(BF10 > 10) with n and ngames ($r_s(23) = 0.74$, $p < 0.001$; $r_s(7) = 0.62$, $p < 0.001$) and negative correlations between P(BF10 < 1/10) with n and ngames ($r_s(23) = -0.49$, $p < 0.001$, $r_s(7) = -0.31$, $p < 0.001$"
I think a stronger rational needs to be given here for mixing Bayesian and frequentist approaches to understanding this problem. These are fundamentally different statistical paradigms with inference performed differently. While they may come to similar conclusions, care needs to be taken here.

Discussion
"n this work, we used BFDA to analyse the effect of sample size, number of games per participant, and effect size on the probability of obtaining significant evidence to support a null or alternative hypothesis,and the probabilities of incorrectly supporting either hypoth-esis**."**
This needs to be contextualized for experiments that are modeled by MAB problems, and not strictly true for all possible experimental designs. I think this is where the real power of this model lies. If one is running experiments with similar designs and can, using prior knowledge, place reasonable predictions around model parameters, then one has a really nice and cheap way of assessing bounds and providing reasonable sample size prediction. However, the moment the experimental design is not MAB-like, then these results may not be strictly true.

"examine latent variables in relation to BFDA."
Again, specific examples of this from results would be extremely beneficial.

"In the example we used an uninformative prior"
This needs to be placed in methods section to orient the reader to what priors are being used. Add

in uniform prior parameters as well. See (Kruschke, 2021).

Works Cited

Blackwell, D.L., Ramamoorthi, R.V., 1982. A Bayes but Not Classically Sufficient Statistic. The Annals of Statistics 10, 1025–1026.
Cronin, B., Stevenson, I.H., Sur, M., Körding, K.P., 2010. Hierarchical Bayesian Modeling and Markov Chain Monte Carlo Sampling for Tuning-Curve Analysis. Journal of Neurophysiology 103, 591–602. https://doi.org/10.1152/jn.00379.2009
Gelman, A., 2006. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis 1. https://doi.org/10.1214/06-BA117A
Gelman, A., Carlin, J., 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci 9, 641–651. https://doi.org/10.1177/1745691614551642
Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models, 1st ed. Cambridge University Press, Cambridge, United Kingdom.
Gelman, A., Rubin, D.B., 1995. Avoiding Model Selection in Bayesian Social Research. Sociological Methodology 25, 165. https://doi.org/10.2307/271064
Gelman, A., Shalizi, C.R., 2013. Philosophy and the practice of Bayesian statistics: *Philosophy and the practice of Bayesian statistics*. Br J Math Stat Psychol 66, 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x
Kruschke, J.K., 2021. Bayesian Analysis Reporting Guidelines. Nat Hum Behav 5, 1282–1291. https://doi.org/10.1038/s41562-021-01177-7
Kruschke, J.K., 2013. Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General 142, 573–603. https://doi.org/10.1037/a0029146
Kruschke, J.K., 2010. What to believe: Bayesian methods for data analysis. Trends in Cognitive Sciences 14, 293–300. https://doi.org/10.1016/j.tics.2010.05.001
Vize, C.E., Phillips, N.L., Miller, J.D., Lynam, D.R., 2024. On the Use and Misuses of Preregistration: A Reply to Klonsky (2024). Assessment 10731911241275256. https://doi.org/10.1177/10731911241275256

References

1. Cronin B, Stevenson IH, Sur M, Körding KP: Hierarchical Bayesian modeling and Markov chain Monte Carlo sampling for tuning-curve analysis.*J Neurophysiol*. 2010; **103** (1): 591-602 PubMed Abstract | Publisher Full Text
2. Gelman A: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*. 2006; **1** (3). Publisher Full Text
3. Gelman A, Carlin J: Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors.*Perspect Psychol Sci*. 2014; **9** (6): 641-51 PubMed Abstract | Publisher Full Text
4. Gelman A, Rubin D: Avoiding Model Selection in Bayesian Social Research. *Sociological Methodology*. 1995; **25**. Publisher Full Text
5. Gelman A, Shalizi CR: Philosophy and the practice of Bayesian statistics.*Br J Math Stat Psychol*. 2013; **66** (1): 8-38 PubMed Abstract | Publisher Full Text
6. Kruschke JK: Bayesian Analysis Reporting Guidelines.*Nat Hum Behav*. 2021; **5** (10): 1282-1291 PubMed Abstract | Publisher Full Text
7. Kruschke JK: Bayesian estimation supersedes the t test.*J Exp Psychol Gen*. 2013; **142** (2): 573-603 PubMed Abstract | Publisher Full Text
8. Kruschke JK: What to believe: Bayesian methods for data analysis.*Trends Cogn Sci*. 2010; **14** (7): 293-300 PubMed Abstract | Publisher Full Text
9. Vize CE, Phillips NL, Miller JD, Lynam DR: On the Use and Misuses of Preregistration: A Reply to

Klonsky (2024).*Assessment*. 2024. 10731911241275256 PubMed Abstract | Publisher Full Text
10. Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models, 1st ed. Cambridge University Press, Cambridge, United Kingdom.
11. Blackwell, D.L., Ramamoorthi, R.V., 1982. A Bayes but Not Classically Sufficient Statistic. The Annals of Statistics 10, 1025–1026.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
No

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: I hold a provisional patent on neural control reinforcement learning methods, USPTO: 18/083490. I am also a consultant for BECATech, LLC for work unrelated to the present data. I confirm that this potential conflict of interest did not affect my ability to write an objective and unbiased review of the article.

Reviewer Expertise: Bayesian Statistics, Neuroscience, Neural Engineering, Reinforcement Learning in Neuroscience

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 25 Apr 2026

Sarah Schreiber

It was a pleasure to read your article detailing Bayes Factor design of experiments and its incorporation into latent variable modeling. Much well deserved interest has recently been received to incorporate Bayesian methods to the design of experiments, and I'm pleased to see your method adding to the conversation. I believe that modeling experiments during the experimental design phase with well-defined models of decision processes can certainly help bound and inform the BFDA process in meaningful ways. However, there are several areas in the

manuscript that require further refinement to enhance both its readability and the broader adoption of your model and methodology.
Response: We are thankful to Reviewer 1 for a generally positive evaluation of our work.

Point 1.
I believe that the description of the MAB task needs much better description. As it stands, it's placed in the context of a pain model that is never described, and the definition of a painful stimulus is vague at best. I also believe that this article often verges into vague descriptions of what is being done. I've tried to highlight sections below, but I believe portions can be rewritten to explicitly state what is being done. This will greatly increase readability.
Response: We thank the reviewer for this feedback. We have revised the description of the multi-armed bandit (MAB) task to improve clarity and specificity. The revised text now provides a clearer definition of the task structure, the exploration-exploitation trade-off, and the nature of the task outcomes. We also explicitly describe the pain-related application of the binary MAB, specifying the delivery versus absence of a painful stimulus (e.g., electrocutaneous stimulation), and we clarify the fixed task horizon and trial structure.
Revised Sections: "To validate the analysis pipelines and to provide a concrete example for the proposed approach, a binary two-armed bandit task was simulated. The MAB is a widely-used paradigm to investigate the exploration-exploitation dilemma in behavioural science (Danwitz *et al.*, 2022; Daw *et al.*, 2006; Gershman, 2019), in which the agent (for example the participant or a reinforcement learning agent) repeatedly chooses between multiple actions or "arms". Over time the agent learns the reward probabilities associated with each arm. To maximise the cumulative outcome across trials, the agent must balance exploring the available arms in order to reduce uncertainty about their outcomes (exploration) and using the information they have gathered so far to choose the optimal option (exploitation)." "The outcomes can be continuous, for example in the form of a scalar monetary reward, where participants learn the reward distribution (e.g. mean and uncertainty), or binary, such as the presence versus absence of a food reward or a negative (aversive) stimulus. The binary version of the MAB task is often used in pain research, where a painful stimulus (e.g., an electrocutaneous shock) is administered. The two possible outcomes would be the delivery of an aversive (e.g. painful) stimulus and its absence (Krypotos *et al.*, 2022a; Krypotos *et al.*, 2022b). In the computational modelling of these outcomes, each outcome must be assigned a numerical value. Here, the outcome of each choice was coded as 0 or -1, where -1 represents an aversive outcome." **"**To facilitate differentiation between exploration and exploitation, the agent has information on both options available to them prior to their first choice, so that the agent possesses information to base their exploitation on (Wilson et al., 2014). This is achieved by presenting the agent with four actions and their immediate results, after which they are free to make their own choices. The task was implemented with a fixed horizon of 10 choices, defining the number of trials within each game, in that each game consisted of four observed trials followed by 6 free choice trials. The outcome probabilities were combinations of the probabilities 0.1, 0.3, and 0.9, with the probability assigned to each arm changing after each game."

Point 2. It should be noted that the use of Bayes factors in model comparisons can be tricky, with the potential to reduce meaningful Bayesian inference to a modified p-value. Many argue that the power of Bayesian thinking lies in using prior knowledge to build the best possible models and to avoid the temptation to just compare every model available. For a good discussion of these

caveats, see (Gelman and Rubin, 1995; Gelman and Shalizi, 2013). Providing a discussion of and context to BFDA may help allay some of these criticisms of BF analyses.
Response: We thank the reviewer for this comment. We have added a discussion that contextualises our use of Bayes factors within the broader Bayesian literature, explicitly acknowledging common concerns about threshold-based interpretations. We clarify that Bayes factors are used here as a pragmatic tool for design analysis, rather than as a replacement for full Bayesian inference.
Revised Section: "It is important to highlight that while this analysis summarizes Bayes factors using common thresholds, their interpretation is not inherently discrete. The primary strength of Bayes factors lies in their continuous nature and in their role within a broader Bayesian framework that emphasizes principled model construction and the use of prior knowledge (Aczel et al., 2020; Coventry & Bartlett, 2024; Gelman & Rubin, 1995; Gelman & Shalizi, 2013). However, in the present work, Bayes factors are used specifically as a pragmatic tool for design analysis, where threshold-based summaries aid decision-making. In retrospective analyses, Bayes factors should still be interpreted within a full Bayesian framework rather than being reduced to a modified p-value."

Point 3.
I also believe this article would benefit greatly from following the Bayesian analysis reporting guidelines (BARG) for article clarity and reproducibility (Kruschke, 2021).
Response*:* We thank the reviewer for this helpful suggestion. We have now incorporated several relevant elements of the Bayesian Analysis Reporting Guidelines for reporting retrospective Bayesian inference to improve the clarity and transparency of our approach for prospective Bayes factor design analysis. Specifically, we now more clearly state the goals of the analysis, explicitly define the observed data and inferred parameters, and clarify the specification and justification of the priors used. We believe these additions improve readability and reproducibility while remaining appropriate to the prospective focus of the present work.
Revised Sections: "The goal of this work is a prospective design analysis, specifically to evaluate how sample size, number of games per participant, and effect size influence Bayes factors and parameter recovery for a latent exploration parameter in a multi-armed bandit task. We first analysed the accuracy of Bayesian parameter estimations of latent variables, within a population, as well as between two groups, using simulated behavioural data in a multi-armed bandit task (Krypotos *et al.*, 2024). In a second step, we combined BFDA with simulations of behavioural data to explore the relationship between sample size and the strength of evidence for both null and alternative hypothesis." **"**To summarise, the primary outcome per game is the number of aversive outcomes (0–6), while the main inferred parameter is the exploration parameter τ. The between group comparison focusses on the difference in the exploration parameter Δμ_τ. For both tau and the difference in population means (Δμ_τ), we used a discrete uniform prior defined over a bounded parameter space ($\tau \in [0.01, 3]$ and $\Delta\mu\_\tau \in [-3, 3]$)."

Specific Comments
Abstract
Point 4.
Interest in the use of Bayesian statistics I believe extends well beyond just the fact that one can incorporate prior knowledge. While that is certainly a powerful component of Bayesian design,

other advantageous features include the fact that inference is performed using the data on hand without implicit assumptions on population distributions, the ability to easily handle data distributions which are not strictly normal, direct quantification of uncertainty under inference, and the ability to iteratively update hypotheses as data is observed.
Response: We thank the reviewer for this comment. Due to abstract word limits, these points could not be fully included. However, we have reworded the Abstract to better reflect the broader advantages of the Bayesian framework and expanded the Introduction to highlight several of the reviewer's suggestions.
Revised Section: Abstract: "Bayesian statistics offers a flexible framework that supports iterative updating of hypotheses and the incorporation of prior information, amongst other advantages." Introduction: "Concurrently, Bayesian methods are gaining recognition for their advantages, which include the incorporation of prior knowledge into statistical processes, the ability to quantify evidence for both null and alternative hypotheses, accommodate non-normal data, and directly represent uncertainty through probability distributions (Jeffreys, 1935; Jeffreys, 1961; Wagenmakers, 2007; Blackwell & Ramamoorthi, 1982;Gelman & Shalizi, 2013; Kruschke, 2010; Kruschke, 2013)."

Introduction:
Point 5.
"*Designing an effective experiment can be challenging due to the number of parameters that need to be considered." While this is true, I believe this is a bit reductive, potentially to the point of trivializing design of experiments (DOE). Rewording of text and adding appropriate discussion of the goals of DOE would add important context to the reader, especially those relatively new to the science and admittedly art of DOE*.
Response: We thank the reviewer for this comment and have revised the Introduction accordingly.
Revised Section: "Designing an effective experiment is a multifaceted process that begins with formulating clear research questions and hypotheses, selecting appropriate methodologies, and aligning design choices with the underlying theoretical framework. Beyond these conceptual considerations, the appropriate analyses and practical parameters must be carefully determined to ensure valid, reliable, and adequately powered statistical inference (Czitrom 2012; Garud et al., 2017; Tůmová et al., 2018).

Point 6.
"To improve a study's effectiveness, it is recommended to conduct a thorough prospective design analysis rather than relying purely on a retrospective approach". While true, I believe it's important to mention preregistration as part of the current initiatives to improve reproducibility. For a thorough discussion of the pros and cons of preregistration, see (Vize et al., 2024).
Response: We thank the reviewer for this suggestion. We have revised the manuscript to explicitly reference preregistration as part of broader reproducibility initiatives.
Revised Section: "Prospective design analyses can help optimize the use of available resources, which has become increasingly important considering recent concerns about 'research waste' (Ioannidis *et al.*, 2014; Macleod *et al.*, 2014; Storz-Pfennig, 2017). These approaches align with efforts to improve reproducibility and replicability, such as preregistration and sharing of code and data (Munafò et al., 2017; Vize et al., 2025). As BFDA requires the primary analysis to be specified a priori, it integrates well with these practices."

Point 7.
"Bayesian methods are gaining recognition for their advantages, as they allow the incorporation of prior knowledge into statistical processes." Again, I think much more can be said about the advantages of Bayesian approaches. See comments above in abstract section, and the following references (Blackwell and Ramamoorthi, 1982; Gelman and Shalizi, 2013; Kruschke, 2013, 2010).
Response: We agree with the reviewer and have revised this section to more clearly emphasize the general advantages of Bayesian statistics and their role in Bayes Factor Design Analysis.
Revised Sections: "Concurrently, Bayesian methods are gaining recognition for their advantages, which include the incorporation of prior knowledge into statistical processes, the ability to quantify evidence for both null and alternative hypotheses, accommodate non-normal data, and directly represent uncertainty through probability distributions (Jeffreys, 1935; Jeffreys, 1961; Wagenmakers, 2007; Blackwell & Ramamoorthi, 1982;Gelman & Shalizi, 2013; Kruschke, 2010; Kruschke, 2013).

Point 8.
"Traditionally, sample size determination methods rely on frequentist statistics, but there has been an ongoing critique among statisticians and methodologists regarding these approaches." While true, I'd argue that the primary goal of DOE is not just to determine sample size, but to determine sample size to ensure an experiment has the statistical power to detect if a desired effect is truly present. Sample size is a means to this end of the question experimenters are truly interested in.
Response: We thank the reviewer for this important clarification. We have revised the Introduction to more clearly emphasize that sample size determination is a means to ensuring adequate statistical power to detect meaningful effects, rather than an end in itself.
Revised Section: "Designing an effective experiment is a multifaceted process that begins with formulating clear research questions and hypotheses, selecting appropriate methodologies, and aligning design choices with the underlying theoretical framework. Beyond these conceptual considerations, the appropriate analyses and practical parameters must be carefully determined to ensure valid, reliable, and adequately powered statistical inference (Czitrom 2012; Garud et al., 2017; Tůmová et al., 2018)." "Traditionally, the focus of predictive analyses has been on determining the sample size required to ensure adequate statistical power to detect meaningful effects, which is a useful step to ensure the quality and validity of the experiment and all conclusions drawn from it. Methodologies for determining sample sizes have long primarily relied on frequentist statistics, but there has been an ongoing critique among statisticians and methodologists regarding these approaches. Among the main concerns are the common misinterpretation of *p*-values and significance testing (Benjamin *et al.*, 2018; Gigerenzer, 2004; Morrison & Henkel, 2017; Wagenmakers, 2007; Wagenmakers *et al.*, 2018)."

Point 9.
"a-priori". Please change to a priori
Response: This has been revised.
Revised Section: "BFDA evaluates the distribution of the Bayes factors for a given experimental design, providing a powerful alternative to frequentist a priori power

analyses."

Point 10.
"BFDA assumes a population with predetermined attributes and conducts simulations on repeated sampling from this population". In the context of a tutorial article, I believe concrete examples are going to be needed to best orient the reader to the goals of BFDA. One way to do this is to provide a quick hypothetical experiment, outlining what example predetermined attributes may be, and how one might conduct simulations. Are there experiments where simulations are generally widely available and accepted? What about cases of a novel experiment, how might one deal with model selection? These are questions that should be addressed to best help the reader adopt this framework.
Response: We appreciate the reviewer's suggestion to include a hypothetical example. However, as the manuscript develops the example of the MAB in detail throughout the remainder of the tutorial, we believe that introducing an additional hypothetical scenario would be beyond the scope of the current article and may create confusion for the reader. Instead, we have revised this section to highlight the goal of BFDA, to clarify the notion of population-level assumptions and elaborated on how distributional properties may be specified in both established and novel paradigms.
Revised Section: "BFDA assumes that the variable of interest is defined at the population level, and that observations are sampled from this population. In well-established paradigms, distributional properties and effect size estimates of the variable at population level can be informed by previous research. In novel paradigms, standard distributions (e.g. the normal distribution) may be assumed and sensitivity analyses across possible effect sizes can be used to assess robustness (Schönbrodt & Wagenmakers, 2018). For each sample generated under this population model, the comparative evidence between the null hypothesis and the alternative hypothesis is measured by calculating the Bayes factor. Repeating this process yields a distribution of Bayes factors, which can be used to evaluate and compare design approaches and thus optimise the experimental setup."

Point 11.
"However, in psychology and neuroscience, we often deal with latent variables that first need to be inferred from the data collected."Absolutely, and I'm glad this is mentioned. It should be noted that in neuroscience and social sciences, initial passes where latent variables are truly hidden is often addressed with hierarchical Bayesian models. See (Cronin et al., 2010; Gelman, 2006; Gelman and Hill, 2006).
Response: We thank the reviewer for calling these references to our attention and have revised the section accordingly.
Revised Section: "However, in psychology and neuroscience, we often deal with latent variables that first need to be inferred from the data collected. Such cases are commonly addressed using computational models (e.g. of behaviour or neural responses), including hierarchical Bayesian models, which allow latent parameters to be estimated while accounting for variability at multiple levels (Cronin et al., 2010; Gelman, 2006; Gelman & Hill, 2006)." Methods.

Point 12.
"The prior distribution P(θ) represents our knowledge about the distribution of θ ahead of data acquisition. The chosen prior can heavily influence the outcome of an analysis and should

therefore be chosen carefully, as it will bias the estimation." True, especially if the observed data is weak, in which case priors dominate. However, some guidance of proper prior choice should be detailed in this article. Potentially adding a prior predictive checks to at least one model to help inform the reader of exactly this condition.
Response: We are grateful for this comment. While an extensive treatment and comparison of prior specification is beyond the scope of this tutorial, we have added references to relevant literature to provide readers with further guidance.
Revised Section: "The chosen prior can heavily influence the outcome of an analysis and should therefore be chosen carefully, as it will bias the estimation. When prior knowledge about the underlying distribution is limited, researchers often employ uninformative or weakly informative priors, which can nevertheless reflect plausible bonds for the parameters. Given the subjective nature of priors, their selection should be transparently reported and sufficiently motivated. In addition, sensitivity analyses can be conducted to evaluate the robustness of results with respect to the chosen prior. For further discussions on priors, refer to further literature such as Gelman (2006), Van Dongen (2006), or Stefan (2019)."

Point 13.
"Once the likelihoods have been calculated for each group, we can compare θ between two groups, for example between patients and healthy controls."
Need to clarify for the reader whether you mean a point estimate of θ or the distribution of θ.
Response: We thank the reviewer for this comment and have clarified that the comparison refers to distributions rather than point estimates.
Revised Section: "Once the likelihoods have been calculated for each group, we can compare the distribution of θ between two groups, for example between patients and healthy controls."

Point 14.
"Figure 1 panel A"
I think this figure has the promise of being very good, but is a bit confusing to read as it stands. As an example, the box "Compute the posterior" is positioned right next to, what I believe is the incorporation of the prior into the likelihood to create the posterior, but it appears based on where the box is positioned that you mean to say P(theta) is the posterior.
Response: We appreciate these observations regarding Figure 1. This figure has now been revised to improve clarity and understanding for the reader. The figure legend has also been revised.

Point 15.
"HDPI is a method of obtaining a credible interval for skewed probability distributions."
Need to describe credible regions, and I believe you need to redo the definition of the HPDI, as it's not just for skewed distributions. I think I understand what you mean by skewed here, but as it stands its ambiguous given that one can easily define HPDIs for non-skewed distributions.
Response: We thank the reviewer for this helpful clarification. We have revised the definition of the HPDI to more accurately describe it and its average length.
Revised Section: "A further analysis evaluates the accuracy of the estimation of the difference between two groups by assessing the estimation error and the average length of the highest probability density interval (HPDI) to aid in determining an optimal sample size.

The HPDI is a type of credible interval, representing a range of values within which the parameter has a certain probability of falling (e.g. 95%). Among all possible intervals, the HPDI is the credible interval that includes the highest probability densities. Therefore, a 95% HDPI would include the range of values with the highest densities and 95% probability. The interval is defined as $C_{1-\alpha} = \{\theta: p(\theta|Y) > p^*\}$ with $\int_{C_{1-\alpha}} p(\theta|Y)d\theta = 1 - \alpha.$ An example of a HPDI is shown in Figure 1d. The length of the HPDI can serve as a measurement of uncertainty, with narrower intervals reflecting more precise estimates. The average length of the posterior's *HPDI* can be calculated in a two-step approach (Joseph & Bélisle, 1997). Let *l'* be the length of the *HPD* $_{1-\alpha}$ interval for a given dataset Y_i . The average length *l** can be calculated from *l'* by multiplying it by the probability of Y_i being the outcome for this measurement and integrating over all possible outcomes *y* = [Y_1, Y_2,...] $l^* = \int l'(Y)p(Y|n)dY$. $p(Y|n)$ is the posterior predictive density and can be computed as $p(Y|n) = \int p(Y|\theta, n)\pi(\theta)d\theta$."

Point 16.
"Figure 1"
Consider reorganizing the panels of Figure 1 as they are referenced out of order in text, making the figure a little harder to read.
Response: This figure has now been revised to improve clarity and understanding for the reader. In-text references to Figure 1 have been revised to ensure that all panels are referred to in order.

Point 17.
"The probability of receiving a certain number of painful stimuli in one game is dependent on the choices of the participant on one hand and the reward probabilities of each arm on the other hand."
So this is verging into confusion. I understand that MAB problems are used in models of pain as mentioned above, but if this terminology is going to be adapted in the method (painful stimuli), then this model needs to be fleshed out. Otherwise, I don't not have a well-defined view of what MAB model parameters, such as reward and painful stimuli mean. I'd suggest either removing the pain model from the MAB context and explicitly defining model rewards and failures or totally rope the description of the MAB problem into the pain model context.
Response: We recognize that combining the MAB framework with a pain-related example may have been unclear. The aversive stimulus was chosen due to its relevance in the pain literature. To improve clarity, we revised the explanation and changed the terminology to consistently distinguish the general MAB model from its experimental instantiation, replacing "painful stimulus" with "aversive outcome" in the modelling context. This is also addressed in the response to point 1.
Revised Sections: "The outcomes can be continuous, for example in the form of a scalar monetary reward, where participants learn the reward distribution (e.g. mean and uncertainty), or binary, such as the presence versus absence of a food reward or a negative

(aversive) stimulus. The binary version of the MAB task is often used in pain research, where a painful stimulus (e.g., an electrocutaneous shock) is administered. The two possible outcomes would be the delivery of an aversive (e.g. painful) stimulus and its absence (Krypotos *et al.*, 2022a; Krypotos *et al.*, 2022b). In the computational modelling of these outcomes, each outcome must be assigned a numerical value. Here, the outcome of each choice was coded as 0 or -1, where -1 represents an aversive outcome." "The binary two-armed bandit task was simulated as described and the number of draws with an aversive outcome was recorded. The probability of receiving a certain number of aversive outcomes in one game is dependent on the choices of the participant on one hand and the reward probabilities of each arm on the other hand."

Point 18.
"The learning rate α was fixed at 0.1"
This seems a bit high of a learning rate value. Why was this chosen and if there is appropriate literature, cite it here.
Response: We thank the reviewer for this comment. The value α = 0.1 was selected as a fixed illustrative parameter to facilitate the simulations. We have adjusted the wording to make the considerations clearer.
Revised Section: "In the present simulations, the learning rate α was arbitrarily fixed at 0.1 for illustrative purposes. Although learning rates may vary, particularly in aversive contexts (Krypotos et al., 2022b; Wang et al., 2018), our conclusions are not contingent on this specific choice."

Point 19.
"Therefore, there are seven possible outcomes, as the agent can receive between zero and six painful stimuli." Again, this needs much better defining. "For each choice, there are four possible combinations of arm chosen and aversive or neutral reward. For six free choice trials there are a total of 46 possible paths, or combinations of choices and outcomes." Moving this earlier in the text will help with understanding of the MAB model.
Response: We thank the reviewer for these helpful comments. We have revised and restructured the entire MAB section, moving the description of possible choice-outcome combinations and clarifying the wording to improve overall comprehensibility.
Results

Point 20.
"The use of the BFDA as a method for prospective design analysis allows us to consider a number of factors. However, it is essential to evaluate the algorithm estimating the latent variable as a preliminary step." This needs to be contextualized for understanding. What factors are you considering? What latent variables are you hoping to describe? What specific algorithm do you mean here? This sentence is quite vague. Being quite specific in what results you are showing will help readability greatly.
Response: We revised this section for clarity, more clearly specifying the goals of the BFDA, the outcome parameters, and the analytical approach.
Revised Section: "The use of the BFDA as a method for prospective design analysis allows us to systematically evaluate design parameters, including sample size, number of games, task-specific features such as number of arms and aversive outcome probabilities, and the planned analysis approach. In our example experiment of the binary MAB with aversive

outcomes, our hypothetical question is to compare the exploration parameter τ between two groups. This is a latent parameter which influences the agents' choices (eq.14). Our analysis consists of three main steps summarized in Figure 1c. First, we verify the estimations of population mean for one population. Next, we compare the population means between two group by examining the distribution of Bayes factors, which indicate whether there is evidence for or against a difference in population mean between groups. This analysis is carried out to determine possible effect sizes and the effect of experimental parameters, such as the number of participants, as well as the number of games per participant on the evidence. In the last step, we analyse potential errors when estimating a difference in means between groups and how they are affected by the experimental parameters (number of participants and number of games per participant)."

Point 21.
"Once the algorithm has been validated, it can be used for the BFDA, in which repeated sampling from a population is simulated and the planned analysis carried out on the sample."
A descriptive flow chart figure would be extremely useful here.
Response: We thank the reviewer for this helpful suggestion. Figure 1c presents a flowchart of the validation and BFDA procedures. We have now adjusted the text to improve clarity.
Revised Section: "In our example experiment of the binary MAB with aversive outcomes, our hypothetical question is to compare the exploration parameter τ between two groups. This is a latent parameter which influences the agents' choices (eq.14). Our analysis consists of three main steps summarized in Figure 1c. First, we verify the estimations of population mean for one population. Next, we compare the population means between two group by examining the distribution of Bayes factors, which indicate whether there is evidence for or against a difference in population mean between groups."

Point 22.
"The individual exploration parameter for each simulated participant was drawn from a normal distribution with population mean μτ". Point to the exact parameter referenced in the methods, as right now this is vague.
Response: We have adjusted the text to refer to the softmax algorithm, where τ was first introduced and to include more information on the population.
Revised Section: "For each simulated participant, the individual exploration parameter τ, as defined in the softmax algorithm (14), was drawn from a normal distribution with population mean μ_{τ} ranging from 0.05 to 1 in increments of 0.05, and a standard deviation of $\sigma = 0.02$. The goal was to then estimate the population mean within the Bayesian framework as described (see Figure 1c, Validation of estimations of population mean)."

Point 23.
"The overall fit of the regression model was statistically significant ($F(3,496) = 136300$, $p < 0.001$, $R2 = 0.999$)" This regression model isn't described in the methods. Please detail what exactly is done here. It's also unclear why a frequentist approach here was used to validate the algorithm, given that the aim was a Bayes factor approach. This needs better detailing for why this was done.
Response: We thank the reviewer for this important point. We have revised the Methods section to explicitly describe the regression model used and clarify its role. We now state that the regression analysis is a descriptive check assessing the correspondence between

the algorithm's estimates and the known generative values in simulation. The analysis is not part of the Bayesian inferential framework and is not intended to validate the Bayes factor approach. The frequentist formulation was used solely for transparency and ease of interpretation in this auxiliary verification step.
Revised Section: "To evaluate the accuracy of this estimation, we compared the point estimates calculated as stated in (5) with the true exploration parameters. This was done using a linear regression model with point estimates as the dependent variable and true mean difference, sample size, and their interaction as predictors. The regression model, as well as the subsequent models, were estimated using ordinary least squares in the seaborn python package (Seabold & Perktold, 2010). These regression analyses and correlations are descriptive summaries of the observed relationships and are not part of the Bayesian inferential framework or the Bayes factor-based design analysis. They are used solely as a descriptive check to assess whether the simulated data contain informative signal regarding the accuracy of the estimated population means. We then calculated the mean squared estimation error (MSEE) and evaluated its relationship with sample size using Spearman's rank-order correlation. We also computed a regression model with the MSEE as the dependent variable and the sample size, true population mean, and their interaction as predictors, allowing for nonlinear relationships. After inspection, the sample size was logarithmically transformed, and the population mean was exponentially transformed. The scaling constants used in these transformations were determined in prior curve-fitting procedures." "Next, we analysed the estimated difference in mean between the two populations by comparing the estimated difference between the two populations  to the true difference $\Delta \mu_{\tau}$. 250 simulations were run for each combination of $\Delta \mu_{\tau}$ ranging from 0 to 1 in increments of 0.05 and sample size per group ranging from 10 to 58 in increments of 2. A linear regression model with point estimates of the difference as the dependent variable and true mean difference, sample size, and their interaction as predictors was used as a descriptive summary of the correspondence between estimated and generative values. Again, these analyses are not part of Bayesian inferential framework. The MSEE was calculated for the estimated difference and its relationship with sample size and mean difference was examined using a regression model with MSEE as the dependent variable and the sample size, true mean difference, and their interaction as predictors, incorporating nonlinear transformations of the predictors. Log transformation was applied to the sample size, and exponential transformation was applied to the difference in population mean, with scaling constants established in earlier curve-fitting operations."

Point 24.
"250 simulations were run for each combination of Δμτ ranging from 0 to 1 in increments of 0.05 and sample size per group ranging from 10 to 60 in increments of 2."
Please make this a little clearer. What simulation? Markov decision process? Is this MAB RL? Better descriptions of what is being done need to be added.
Response: We thank the reviewer for this comment. We have expanded the description of the simulation procedure to clarify that each replication involved simulating participant choices within MAB reinforcement learning framework, followed by computation of the Bayes factor as part of the BFDA. The revised text now more explicitly outlines the sequential steps of the simulation and analysis process.
Revised Section: "To evaluate the effects of sample size and difference in means on the

evidence for a difference between groups, we performed the BFDA. First, we drew random samples of the exploration parameter τ from the two populations with different means, with each sampled value representing one simulated participant. Behavioural outcomes within the binary MAB were then generated based on τ, and the distribution $p(Y \mid \Delta\mu\_\tau)$ was computed as described in Equation 6, from which the Bayes factor was calculated (Equation 8). This procedure was repeated 250 times for each combination of the sample size of each group ranging from 10 to 58 in increments of 2 and difference in means ranging from 0 to 0.95 in increments of 0.05. To determine the population means for a given mean difference, the mean of the first population was randomly sampled from a uniform distribution between 0 and $1 - \Delta\mu\_\tau$, and the mean of the second population was defined by adding the specified difference. The Bayes factor was calculated for each of the 250 simulations, yielding a distribution of Bayes factors reflecting the evidence for or against a difference in means at each combination of sample size and mean difference."

Point 25.
"Figure 2"
This is interestingly. What is happening at 0.05? It seems that BF dominate predictions except in this small window. When there's 0 difference in the mean, Bayes is doing create, but minimal and not moderate uncertainty proves problematic?
Response: We thank the reviewer for this observation. Because the null hypothesis is specified as a point hypothesis (i.e., Δμτ = 0), even extremely small deviations from zero imply that the null is technically false. In this region of very small effect sizes, the Bayes factor may not strongly favour the alternative unless the sample size is sufficiently large. As a result, moderate evidence can appear unstable in this narrow range, reflecting the difficulty of distinguishing very small effects from zero. We have explicitly added this fact to the text.
Revised Section: "With an increase of $\Delta\mu_\tau$ and *n* the Bayes factors are more likely to support the correct hypothesis and the stronger the evidence in favour of the correct hypothesis. When the difference in means is zero, the evidence correctly supports the null hypothesis. However, even very small deviations from zero technically contradict the null hypothesis, which can challenge the algorithm. Detecting and strongly supporting such minimal differences requires very large sample sizes.**"**

Point 26.
"For Δμτ ≠ 0 this probability is the probability of wrongfully accepting the null hypothesis, which corresponds to the type II error."
In Bayesian models, type I/II errors are generally not considered for a variety of reasons related to how Bayesian inference is performed. Bayesian approaches instead talk of Type M or Type S errors generally. You can derive a type I/II error for Bayesian approaches, but care must be taken in doing so. See (Gelman and Carlin, 2014).
Response: We thank the reviewer for this comment and have revised the relevant sections to explicitly define and label Type S and Type M errors.
Revised Sections: "Similarly, we can calculate the probability of receiving a Bayes factor BF_{10} smaller than 1/3 which would indicate at least moderate evidence in favour of the null hypothesis (Figure 3b). When this occurs despite a true nonzero mean difference, it constitutes a Type S (sign) error in the Bayesian framework." "To gain an understanding of

how the Bayes factor relates to the true estimate of the difference in mean, we can determine the magnitude error for those simulations that result in a Bayes factor greater than 10. The magnitude error (Type M error) is calculated by dividing the difference between the estimated value and the true difference in mean by the true difference in mean. This reflects an overestimation of the true effect size in statistically significant results (Gelman & Carlin, 2014)."

Point 27.
"Figure 3"
Please label color map axis in this figure.
Response: The colour map axis has been labelled in Figure 3 as requested. This change has also been applied to Figure 4.

Point 28.
"This is supported by positive correlations between P(BF10 > 10) with n and ngames (rs(23) = 0.74, p < 0.001;rs(7) = 0.62, p < 0.001) and negative correlations between P(BF10 < 1/10) with n and ngames (rs(23) = –0.49, p < 0.001, rs(7) = –0.31, p < 0.001"
I think a stronger rational needs to be given here for mixing Bayesian and frequentist approaches to understanding this problem. These are fundamentally different statistical paradigms with inference performed differently. While they may come to similar conclusions, care needs to be taken here.
Response: We thank the reviewer for raising this important point. As in Point 23, the reported correlations are not intended as inferential statistics but as descriptive summaries of simulation outcomes. We have clarified this distinction in the manuscript to explicitly separate these summaries from the Bayesian inferential framework.
Revised Section: "These regression analyses and correlations are descriptive summaries of the observed relationships and are not part of the Bayesian inferential framework or the Bayes factor-based design analysis. They are used solely as a descriptive check to assess whether the simulated data contain informative signal regarding the accuracy of the estimated population means."

Discussion
Point 29.
"In this work, we used BFDA to analyse the effect of sample size, number of games per participant, and effect size on the probability of obtaining significant evidence to support a null or alternative hypothesis, and the probabilities of incorrectly supporting either hypothesis."
This needs to be contextualized for experiments that are modeled by MAB problems, and not strictly true for all possible experimental designs. I think this is where the real power of this model lies. If one is running experiments with similar designs and can, using prior knowledge, place reasonable predictions around model parameters, then one has a really nice and cheap way of assessing bounds and providing reasonable sample size prediction. However, the moment the experimental design is not MAB-like, then these results may not be strictly true.
Response: We thank the reviewer for this insightful comment. We agree that the present findings are specific to experiments modelled within a multi-armed bandit framework. We have revised the manuscript to more clearly contextualise the results within this task structure and to avoid overgeneralization.
Revised Section: "In this work, we used BFDA in the example of a binary multi-armed bandit

(MAB) task with aversive outcomes to analyse the effect of sample size, number of games per participant, and effect size on the probability of obtaining significant evidence to support a null or alternative hypothesis, and the probabilities of incorrectly supporting either hypothesis." "It is important to note that these results are specific to the task used. Although some findings may generalize to similar MAB learning tasks, they do not necessarily extend to other experimental paradigms."

Point 30.
"examine latent variables in relation to BFDA."
Again, specific examples of this from results would be extremely beneficial.
Response: We thank the reviewer for this comment. We have revised the text accordingly to improve clarity.
Revised Sections: "The example of a multi-armed bandit (MAB) task with prior information was used to examine latent variables in relation to BFDA. Specifically, we considered a hypothetical experiment designed to assess differences in the exploration parameter (τ), estimated via the softmax decision-rule, between two groups. We showed how sample size, number of games per participant and effect size influence the probability of obtaining evidence for group differences in τ." "We found that it can be challenging to determine the difference in means between two groups. Large sample sizes are needed for a strong probability of detecting evidence in favour of difference, where present. If there is strong evidence supporting the alternative hypothesis, the probability of overestimating the difference in means is relatively high. In the example we used an uninformative prior, but depending on the experiment and existing literature, this should be adjusted. A well-informed prior can lead to higher probabilities of detecting a difference between groups."

Point 31.
"In the example we used an uninformative prior"
This needs to be placed in methods section to orient the reader to what priors are being used. Add in uniform prior parameters as well. See (Kruschke, 2021).
Response: We thank the reviewer for pointing this out. We have revised the methods section to explicitly state the prior specification, including the parameters of the uniform prior.
Revised Section: "For both tau and the difference in population means (Δμ_τ), we used a discrete uniform prior defined over a bounded parameter space ($\tau \in [0.01, 3]$ and $\Delta\mu\_\tau \in [-3, 3]$)."

References 1. Cronin B, Stevenson IH, Sur M, et al.: Hierarchical Bayesian modeling and Markov chain Monte Carlo sampling for tuning-curve analysis. *J Neurophysiol.* 2010; **103** (1): 591-602 PubMed Abstract | Publisher Full Text 2. Gelman A: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis.* 2006; **1** (3): | Publisher Full Text 3. Gelman A, Carlin J: Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspect Psychol Sci.* 2014; **9** (6): 641-51 PubMed Abstract | Publisher Full Text 4. Gelman A, Rubin D: Avoiding Model Selection in Bayesian Social Research. *Sociological Methodology.* 1995; **25** | Publisher Full Text 5. Gelman A, Shalizi CR: Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol.* 2013; **66** (1): 8-38 PubMed Abstract | Publisher Full Text 6. Kruschke JK: Bayesian Analysis Reporting Guidelines. *Nat Hum Behav.* 2021; **5** (10): 1282-1291 PubMed

Abstract | Publisher Full Text 7. Kruschke JK: Bayesian estimation supersedes the t test. *J Exp Psychol Gen.* 2013; **142** (2): 573-603 PubMed Abstract | Publisher Full Text 8. Kruschke JK: What to believe: Bayesian methods for data analysis. *Trends Cogn Sci.* 2010; **14** (7): 293-300 PubMed Abstract | Publisher Full Text 9. Vize CE, Phillips NL, Miller JD, et al.: On the Use and Misuses of Preregistration: A Reply to Klonsky (2024). *Assessment.* 2024; 10731911241275256 PubMed Abstract | Publisher Full Text Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models, 1st ed. Cambridge University Press, Cambridge, United Kingdom. Blackwell, D.L., Ramamoorthi, R.V., 1982. A Bayes but Not Classically Sufficient Statistic. The Annals of Statistics 10, 1025–1026.

Competing Interests: No competing interests were disclosed.