Bayesian paired comparison with the bpcs package

Citation for the original published paper (version of record):
Issa Mattos, D., Martins Silva Ramos, É. (2021)
Bayesian paired comparison with the bpcs package
Behavior Research Methods & Instrumentation, In Press
http://dx.doi.org/10.3758/s13428-021-01714-2

Behavior Research Methods
https://doi.org/10.3758/s13428-021-01714-2

Bayesian paired comparison with the *bpcs* package

David Issa Mattos[1] · **Érika Martins Silva Ramos**[2]

Accepted: 16 September 2021

Abstract

This article introduces the `bpcs` R package (Bayesian Paired Comparison in Stan) and the statistical models implemented in the package. This package aims to facilitate the use of Bayesian models for paired comparison data in behavioral research. Bayesian analysis of paired comparison data allows parameter estimation even in conditions where the maximum likelihood does not exist, allows easy extension of paired comparison models, provides straightforward interpretation of the results with credible intervals, has better control of type I error, has more robust evidence towards the null hypothesis, allows propagation of uncertainties, includes prior information, and performs well when handling models with many parameters and latent variables. The `bpcs` package provides a consistent interface for R users and several functions to evaluate the posterior distribution of all parameters to estimate the posterior distribution of any contest between items and to obtain the posterior distribution of the ranks. Three reanalyses of recent studies that used the frequentist Bradley–Terry model are presented. These reanalyses are conducted with the Bayesian models of the `bpcs` package, and all the code used to fit the models, generate the figures, and the tables are available in the online appendix.

Keywords Bayesian paired comparison · Bradley-Terry · Davidson

Introduction

Paired comparison data analysis arises in several contexts, such as selecting preferences or ranking items (Cattelan, 2012). For example, a person might be presented with questions such as "Which brand of pizza do you prefer?" and needs to choose between pairs, such as "Tombstone or DiGiorno?" or "DiGiorno or Freschetta?" (Luckett, Burns, & Jenkinson, 2020). The most common modeling technique for paired comparisons and the focus of this article is the Bradley–Terry model and its extensions (Bradley & Terry, 1952).

Ordinal scales can also be used to assess preferences, but they may lead to several difficulties. For example, participants may struggle to use the scale correctly (Coetzee & Taylor, 1996; Petrou, 2003), they may try to self-monitor their answers (Kreitchmann, Abad, Ponsoda, Nieto, & Morillo, 2019; Hontangas et al., 2015), and for specific samples, the scales may not even be effective, such as in animal behavior studies, studies with young children, people with low literacy, or when respondents are using their second language to answer the scales (Luckett et al., 2020; Hopper, Egelkamp, Fidino, & Ross, 2019; Huskisson, Jacobson, Egelkamp, Ross, & Hopper, 2020).

Paired comparisons (also called forced-choice assessments) may be stimulus-centric or person-centric. Psychometric research has developed many model applications for behavioral choice theories that consider the individual differences in psychological attributes (person-centric). These models employ item response theory (IRT) methods, such as the multidimensional pairwise comparison (MPC) (Wang, Qiu, Chen, Ro, & Jin, 2017) and the multi-unidimensional pairwise preference two-parameter logistic (MUPP-2PL) model (Morillo et al., 2016). This line of research focuses on person-centric assessments, which have been vastly used in personality and attitudinal research (Brown, 2016).

In the context of this article and the context of applied psychological and behavioral research, the stimulus-centric approach is the one that fits better, and it is in this context that the Bradley–Terry model could be most useful. Examples are the preferences of natural landscapes (Hägerhäll et al., 2018), preferences for stimulus (Chien

✉ David Issa Mattos
issamattos.david@gmail.com

Érika Martins Silva Ramos
erika.ramos@psy.gu.se

[1] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden

[2] Department of Psychology, University of Gothenburg, Gothenburg, Sweden

Published online: 30 November 2021

et al., 2012), for food (Luckett, Burns, & Jenkinson, 2020; Coetzee & Taylor, 1996), analysis of animal dominance (Abalos, de Lanuza, Carazo, & Font, 2016; Bush, Quinn, Balreira, & Johnson, 2016; Miller et al. 2017)and in the use of pharmacological medication (Meid et al. 2016).

Many models have been developed to extend the Bradley–Terry model. For example, to include ties in the comparisons (Davidson, 1970), to address the problem of ordering items' presentation (Davidson & Beaver, 1977), to compensate for dependency on the data and subject-specific predictors (Böckenholt, 2001), or to add explanatory variables to the items (Springall, 1973).

Although efforts have been made in statistical computing to provide more accurate standard errors and *p*-value estimates for the analyses with the Bradley–Terry model (Turner & Firth, 2012; Cattelan, 2012), most of these works have been focusing on providing software for the frequentist Bradley–Terry model based on the maximum likelihood (Turner & Firth, 2012; Hatzinger & Dittrich, 2012; Turner & Firth, 2020). There are no comprehensive software packages that implement a Bayesian version of the Bradley–Terry model and its many extensions.

Therefore, this article proposes a Bayesian perspective to work with the Bradley–Terry model. Bayesian data analysis can provide advantages to frequentist estimation of paired comparison data. For example, it can provide parameter estimates without the need of specifying a maximum likelihood (allowing to incorporate extensions easily) and in problems where the maximum likelihood does not exist or leads to undetermined probabilities (Ford, 1957; Phelan & Whelan, 2017; Butler & Whelan, 2004). Additionally, it provides better control of type I error (Kelter, 2020), provides more robust evidence towards the null hypothesis (Kruschke, 2013), handles models with many parameters and latent variables (Kucukelbir, Ranganath, Gelman, & Blei, 2015; Carpenter et al., 2017), and allows a probabilistic interpretation of parameter intervals, as opposed to the repeated sampling interpretation (McElreath, 2020; Kruschke, 2013). Specifying the priors for the parameters enables users to incorporate prior knowledge into the model and obtain stricter credible intervals, especially in the ties and order effects extensions discussed in the next section.

This paper introduces the `bpcs` R package (Bayesian Paired Comparison in Stan) and the statistical models implemented in the package. This package aims to facilitate the use of Bayesian models for paired comparison data in research. The `bpcs` package provides a consistent interface for R users to work with these models and several functions for researchers to evaluate the posterior distribution of all parameters, to estimate the posterior distribution of any contest between items, and to obtain the posterior distribution of the ranks. The paper provides Bayesian reanalyses of three recent studies that used the frequentist Bradley–Terry model. These reanalyses are conducted with the Bayesian models of the `bpcs` package, and all the code used to fit the models, and generate the figures, and the tables are available in the online appendix.[1]

Related software

This section includes relevant software for the Bayesian estimation of the Bradley–Terry and Thurstone models. Johnson and Kuhn (2013) provide a mathematical description, code, and discussion about the implementation of Bayesian Thurstonian models for ranking data using the Gibbs Sampler JAGS. However, it is still up to the user of the software to process this data and generate relevant statistics and plots. Gibbs sampling is less efficient than the No-U-Turn Hamiltonian Monte Carlo (NUTS), used in the `bpcs` package. The NUTS is more efficient in terms of effective samples, in cases with higher autocorrelation and in hierarchical models (Nishio & Arakawa, 2019; Carpenter et al., 2017; Hoffman & Gelman, 2014).

Caron and Doucet (2012) proposes a specific Gibbs sampler for the generalized Bradley–Terry model. The proposed approach shows that the Monte Carlo-based samplers are efficient in estimating the parameters. However, their approach is limited to the generalized model and does extend to the extensions provided in the `bpcs` package.

The `sport` R package provides statistical models for sequential paired comparison data (Sport, 2020), such as the Glicko and Glicko2 (Glickman, 2001) models and the Bayesian Bradley–Terry model. The Glicko and Glicko2 models are sequential, and the results depend on the order in which the items are presented (Glickman, 2001; Luckett et al., 2020). The package provides only the simple Bayesian Bradley–Terry model with a Bayesian Approximation Method. However, it does not allow to set up priors or obtain the posterior distribution.

The `pcFactorStan` (Pritikin, 2020) R package implements a Bayesian item response theory model for paired comparison. The items are measured with a Bayesian Bradley–Terry model and the worth values of the contestants are expanded with a latent factor score. While the `pcFactorStan` package can be used to create rank with the Bradley–Terry model, it focuses on answering factor analysis problems. Moreover, it does not provide the extensions for the generalized, hierarchical, or model with ties.

The `thurstonianIRT` R package allows users to fit item response theory models for forced-choice questionnaires. The package implements the models proposed by Brown and Maydeu-Olivares (2011) and Brown (2016). The

[1] https://davidissamattos.github.io/bpcs-online-appendix/

software utilizes a different backend for model estimation (MPlus, laavan, and Stan). While the package does not provide this functionality directly, although the Stan backend can provide the posterior distributions of the parameters of the Bayesian model, this functionality is not provided in the software package.

The `PLMIX` implement finite mixture of Plackett–Luce models. In the case of partial rankings of two items (forced choice assessment) the Plackett–Luce model reduces to the Bradley–Terry model- (Turner, van Etten, Firth, and Kosmidis, 2020). The package focuses on providing point estimates of the Bayesian estimation through Gibbs sampling. Apart from handling ties, the package does not provide the additional extensions of the `bpcs` package.

Corff, Lerasle, and Vernet (2018) provides a custom non-parametric Gibbs sampling MCMC algorithm to approximate the posterior distribution of a Bayesian Bradley–Terry model in the random environment extension. While this extension is not considered in the `bpcs` package, the algorithm does not support the other use cases from the `bpcs` or provide ready to use functions to support applied behavior research.

Seymour et al. (2020) provide a Bayesian application of the Bradley–Terry model to spatial and geographical applications. The proposed extension introduces a multivariate normal prior distribution to model the spatial structure instead of linear regression methods of generalized methods. The `bpcs` package utilizes a linear regression approach to include covariates. While it might not be suitable for the specific spatial application of Seymour et al. (2020), the linear model is the most common model and has been successfully used in applied research when including stimuli covariates (Dittrich, Hatzinger, & Katzenbeisser, 1998; Fleischhaker, 2019; Giambona & Grassini, 2020).

The `bpcs` package utilizes the No-U-Turn sampler implemented in Stan, provides an easy-to-use interface, and a higher number of extensions, such as models with ties, generalized models, subject-specific predictors and hierarchical models, functions to create tables, and plots that facilitate interpretation of the models. Compared to existing software packages, these features offer a higher flexibility, such as combining multiple extensions together, sampling efficiency using the No-U-Turn Hamiltonian Monte Carlo Sampler, and the easiness of use by providing a single consistent interface that hides the mathematical complexity of the models.

Statistical models for paired comparison

The mathematical models presented in this section have their origin in the Thurstone Law of Comparative Judgment (Thurstone, 1927). This law can assess the difference between two stimuli measured by a scale. Thurstone (1927) proposes a series of cases in which assumptions are made to simplify the problem in terms of tractability (Tsukida & Gupta, 2011). In its general form (Case I), the estimation of the difference between two stimuli requires the estimation (or knowledge) of the dispersion of all stimuli and all of its correlations. The most known simplification (Case V) assumes that all options have equal dispersion and are uncorrelated. Due to its computational tractability, Case V has become more popular than the other less-restrictive cases (Cattelan, 2012; Shah et al., 2015). Case V is also referred to as the Thurstone–Mosteller model, and often just as the Thurstonian model (Cattelan, 2012; Handley, 2001; Johnson & Kuhn, 2013).

The Bradley–Terry model (Bradley & Terry, 1952) provides a similar formulation to the Thurstone–Mosteller model, but assumes that the difference between two stimuli is a logistic random variable instead of a normally distributed variable (Cattelan, 2012). In practice, both models yield nearly identical estimates and expected probabilities of one player beating the other (Handley, 2001). Unlike the Thurstone–Mosteller model, the Bradley–Terry model introduced an additional computational simplicity since the logit function has a closed-form expression. Since its introduction in 1962, the Bradley-Terry model has been extended to address a wider range of problems in both the frequentist and Bayesian frameworks. Examples of such problems are the presence of order effects, random effects, ties, subject predictors, generalized models among others.

The `bpcs` implements the Bayesian extensions of the Bradley–Terry model. The choice for this family of models (instead of Thurstone Case I-V) is due to the wide use of Bradley–Terry models in behavior research, the large number of available extensions proposed in research, nearly identical estimates as the Thurstone Case V, and with better computational tractability.

This section provides an overview of the terminology, introduces the simplest case of the Bayesian Bradley–Terry model implemented in the package followed by a discussion of the different extensions.

Terminology

Different research areas work with different terminology and therefore it is worth having further clarifications:

- Players or contestants are synonymous with the items being compared, i.e., the choices of some type of stimuli such as images, sounds, or objects. The `bpcs` package and this article utilize the term player.
- Subjects, participants, or judges are synonymous for the respondents of a questionnaire, or the subject that is selecting between the paired comparison. Apart from

the term multiple judgment sampling, which refers to when a subject judges multiple items, the `bpcs` package and this article utilize the term subjects.
- A contest refers to a single comparison between two players made by a subject.
- Ties or draws refer to the case where a subject does not express a preference for a player in a contest. For example, in a questionnaire that asks subjects to select between two stimuli (the players) a tie could be an option such as "I do not have a preference". To avoid confusion withdrawing samples of a posterior distribution, the `bpcs` package and this article utilize the term ties.

The mathematical models utilize the following basic notation. Additional symbols and notations are presented with the extension that introduces them.

- $\alpha_i > 0$: a latent variable that represents the ability of player i.
- $\lambda_i = \log(\alpha_i)$: the log of the ability of player i.
- $y_{i,j,n}$: the binary result of the contest n between any two players i and j. If player i wins, $y_{i,j,n} = 1$. If player j wins $y_{i,j,n} = 0$.
- $\text{tie}_{i,j,n}$: a binary variable representing if the result of a single contest (at position n) between two players (i and j) was a tie. It assumes $\text{tie}_n = 1$ if it was a tie and $\text{tie}_k = 0$ if it was not a tie.
- N: the number of players.
- σ^2_λ: the variance of the normal prior distribution for the variable λ. A similar notation is used for the prior distribution of the parameters γ, ν, β and U.

To simplify the notation of the presented models below, the index n is omitted.

The Bradley–Terry model

The Bradley–Terry model (Bradley & Terry, 1952) presents a way to calculate the probability of one player beating the other player in a contest. This probability is represented by:

$$\mathcal{P}[i \text{ beats } j] = \frac{\alpha_i}{\alpha_i + \alpha_j} \tag{1}$$

This model is commonly parameterized by the log of the ability variable:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp \lambda_i}{\exp \lambda_i + \exp \lambda_j} \tag{2}$$

This transformation has the benefits of allowing the estimation of a parameter $\lambda \in (-\infty, \infty)$ and simplifying the estimation of the parameters for the frequentist setting with a generalized linear model with the logit function (Cattelan, 2012):

$$\text{logit}(\mathcal{P}[i \text{ beats } j]) = \lambda_i - \lambda_j \tag{3}$$

However, the parameters λ are not uniquely identified and require another constraint, commonly:

$$\sum_{i=1}^{N} \lambda_i = 0 \text{ or } \sum_{i=1}^{N} \lambda_i = 1$$

The first work to propose a Bayesian formulation of the Bradley–Terry model is attributed to Davidson and Solomon (Davidson & Solomon, 1973). They proposed a version of the Bradley–Terry model utilizing a conjugated family of priors and estimators to calculate the posterior distribution of the log abilities of the parameters and the rank of the players. Leonard (1977) discusses the issue of the flexibility an interpretation of the conjugated priors proposed by Davidson and Solomon in the presence of additional explanatory variables and other extensions.

Leonard (1977) suggests moving away from the convention of using conjugated priors and utilizing normal prior distributions for the parameters. The usage of non-conjugated prior distributions has many advantages, including adaptability to extensions, being able to reason and fully specify the prior parameters, ability to extend to hierarchical models, and to specify other prior distribution families.

The two main disadvantages of the approach proposed by Leonard (1977) are the use of approximation methods for the posterior distribution and the computational time. However, the advances in Bayesian computational packages and Markov chain Monte Carlo (MCMC) samplers significantly minimize such disadvantages. The `bpcs` package utilizes normal priors for the λ parameters and models the outcome variable $y_{i,j}$ of a single contest between players i and j with a Bernoulli distribution, based on the probability of winning for $\mathcal{P}[i \text{ beats } j]$.

Therefore, the simple Bayesian Bradley–Terry model can be represented as:

Likelihood:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp \lambda_i}{\exp \lambda_i + \exp \lambda_j} \tag{4}$$

$$y_{i,j} \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j]) \tag{5}$$

Priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma^2_\lambda) \tag{6}$$

The mean of the prior distribution changes the location of the parameters without impacting the relative probability of one player beating the other. Since it does not impact the estimation of the λ_i, its value is set to zero. The standard deviation in the prior distribution represents the space where the model should look for the relative differences in the probabilities. Smaller values for the standard deviations indicate that the relative preferences are close and have many overlaps. Higher standard deviation indicates that the sampler can look for solutions that are very far apart (probabilities closer to 0 or 1). Choosing high standard

deviations for the prior can increase the time to find a solution and possibly divergence between the chains in the sampler. Very low standard deviations are informative priors and imply that the strength parameters are very close to each other. Both the mean and the standard deviation of the prior act as soft constraint, making the model identifiable (Pritikin, 2020; Stan Development Team, 2016).

The parameters λ_i can be used to rank the different players. However, in the Bayesian framework, a single measure is not obtained but rather a posterior distribution of the parameters λ_i. By sampling from the posterior distribution of the log-abilities of the players, it is possible to create a posterior distribution of the ranks of the players, which helps to evaluate the uncertainty in the ranking system.

Davidson model

The first extension to be added to the Bradley–Terry model is the ability of handling ties in the contest between two players. This approach was proposed by Davidson (1970), which adds an additional parameter ν and computes two probabilities: the probability of i beating j given that it was not a tie $\mathcal{P}[i \text{ beats } j|\text{not tie}]$ and the probability of the result being a tie $\mathcal{P}[i \text{ ties } j]$. A Bayesian formulation of the model is represented below:

Likelihood:

$$\mathcal{P}[i \text{ beats } j|\text{not tie}] = \frac{\exp\lambda_i}{\exp\lambda_i + \exp\lambda_j + \exp\left(\nu + \frac{\lambda_i+\lambda_j}{2}\right)} \tag{7}$$

$$\mathcal{P}[i \text{ ties } j] = \frac{\exp\left(\nu + \frac{\lambda_i+\lambda_j}{2}\right)}{\exp\lambda_i + \exp\lambda_j + \exp\left(\nu + \frac{\lambda_i+\lambda_j}{2}\right)} \tag{8}$$

$$y_i \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j|\text{not tie}]) \tag{9}$$

$$\text{tie}_{i,j} \sim \text{Bernoulli}(\mathcal{P}[i \text{ ties } j]) \tag{10}$$

Priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2) \tag{11}$$

$$\nu \sim \mathcal{N}(0, \sigma_\nu^2) \tag{12}$$

The ν parameter in (the log scale) balances the probability of ties against the probability of not having ties. If $\nu \to -\infty$ than $\mathcal{P}[i \text{ ties } j] \to 0$ regardless of the players' abilities, which is equivalent to the Bradley–Terry model. If $\nu \to +\infty$ then $\mathcal{P}[i \text{ ties } j] \to 1$ regardless of the players' abilities. If $\nu = 0$ than $\mathcal{P}[i \text{ ties } j] = \frac{\exp\frac{\lambda_i+\lambda_j}{2}}{\exp\lambda_i+\exp\lambda_j+\exp\frac{\lambda_i+\lambda_j}{2}}$, which means that the probability of ties depends only on the player's abilities. In this last case, if players i and j have equal abilities, they have an equal chance of having a tie, i winning or j winning.

For the Bayesian version of the Davidson model, the choice of prior in the ν parameter refers to the prior belief on how ties are affected by the relative difference in players' abilities. If the intended goal is also to investigate if the probability of wins and ties depend only on the abilities, the mean parameter of the normal prior distribution for ν is set to zero (as in the presented models).

Although the subsequent models are presented only in the Bradley–Terry variation, they can be easily extended and implemented as a Davidson model to handle ties.

Models with order effect

In paired-comparison problems, a problem that can arise is that the order in which two players are presented can lead to a bias in the choice of the comparison. For example, when two items are presented to be chosen, subjects might have a preference for items placed on the left side. Another example that arises from sports competitions is the presence of home-field advantage; athletes (the players in paired comparison) competing in their home field can have an advantage compared to the visitor player.

The order effect can be modeled as either additive or multiplicative. Davidson and Beaver (1977) discuss some of the advantages of the multiplicative model compared to the additive. One important advantage of the Bayesian version is that the value of the order-effect parameter does not depend on the abilities parameters. Therefore, setting the prior distribution of the order-effect parameter is independent of the soft constraints applied to the log-abilities parameter. Additionally, the multiplicative model has easily been introduced to both the Bradley–Terry and the Davidson models when estimating the parameters in the log scale.

To compensate for the order effect using a multiplicative model, Davidson and Beaver (1977) introduced an additional parameter γ. This multiplicative parameter becomes in the log-scale an additive term, as shown below. The Bayesian Bradley–Terry model with order effect can be represented as:

Likelihood:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp\lambda_i}{\exp\lambda_i + \exp\left(\lambda_j + \gamma\right)} \tag{13}$$

$$y_i \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j]) \tag{14}$$

Priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2) \tag{15}$$

$$\gamma \sim \mathcal{N}(0, \sigma_\gamma^2) \tag{16}$$

The γ parameter in (the log scale) reflects the impact of the order effect. If $\gamma \to -\infty$ then $\mathcal{P}[i \text{ beats } j] \to 1$, which means that regardless of the players' abilities, the player i will have an order-effect advantage and will always win the contest. Analogously, if $\gamma \to +\infty$ then $\mathcal{P}[i \text{ beats } j] \to 0$ and player i will have an order-effect disadvantage and will

always lose the contest. If $\gamma = 0$, then player i will neither have an advantage or disadvantage with the order effect, and the probability of wins or ties depends only on the players' abilities and the tie parameter ν. The choice of prior in the γ parameter refers to our prior belief on the location and spread of the value of order effect. If the intended goal is also to investigate if there is an order effect or not, the mean parameter of the normal prior distribution for γ is set to zero (as in the presented models).

Generalized models

Many research problems require the investigation of the effect of players properties in the probability of winning a contest. The extension proposed by Springall (1973) is analogue to the multiple regression case. This extension proposes the use of K predictors that are characteristic of the players (and not of the subjects which is discussed later). $X_{i,k}$ is the k predictor value of player i and β_k is the coefficient of the predictor that we are estimating. Note that for these generalized models, the intercept is not identifiable and therefore not included (Springall, 1973; Stern, 2011). The Bayesian generalized Bradley–Terry can be represented as:

Likelihood:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp \lambda_i}{\exp \lambda_i + \exp \lambda_j} \tag{17}$$

$$\lambda_i = \sum_k^N X_{i,k}\beta_k \tag{18}$$

$$y_i \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j]) \tag{19}$$

Priors:

$$\beta_k \sim \mathcal{N}(0, \sigma_\beta^2) \tag{20}$$

The generalized version of the Bradley–Terry model estimates the parameters β_k. The parameter λ_i is then estimated by the linear model $\lambda_i = \sum_k^N X_{i,k}\beta_k$. The choice of prior in the β parameter refers to the prior belief on the value of the coefficient of each predictor. The presented generalized models utilize the same prior for all β coefficients. Therefore it requires that the values for every k in the $X_{i,k}$ be on the same range, otherwise, the model would have a strong informative prior belief in some coefficients and a weakly informative prior belief in other coefficients. If the predictors' input values $X_{i,k}$ are normalized for every k the larger the coefficient β_k, the higher the influence of that predictor in the probability of winning.

Introducing random-effects to model-dependent data

It is common in the research context to have the same subject to make multiple comparisons, the multiple judgment sampling problem (Cattelan, 2012). A more realistic analysis of the Bradley–Terry model would assume that the comparisons made by the same person are dependent. One approach to address the multiple judgment sampling problems is through the usage of mixed-effects or hierarchical paired-comparison models (Böckenholt, 2001).

Böckenholt (2001) decomposed the paired-comparison model into fixed and a random effect components. The random effects component estimates the subject variation (given S subjects) in each item, while the fixed effect component estimates the average log ability of the player. The random effects term is represented by $U_{i,s}$, where i refers to the player being judged and s to the subject. The Bayesian Bradley–Terry model with random effects can be represented as:

Likelihood:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp \lambda_{i,s}}{\exp \lambda_{i,s} + \exp \lambda_{j,s}} \tag{21}$$

$$\lambda_{i,s} = \lambda_i + U_{i,s} \tag{22}$$

$$y_{i,j} \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j]) \tag{23}$$

Priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2) \tag{24}$$

$$U_{i,s} \sim \mathcal{N}(0, U_{\text{std}}^2) \tag{25}$$

$$U_{\text{std}} \sim \text{Half-}\mathcal{N}(0, \sigma_U^2) \tag{26}$$

This model aims at estimating the parameter U_{std} that represents the standard deviation in the random effects and the difference between subjects. In the Bayesian context, with the Stan probabilistic programming language, it is also possible to estimate the parameters $U_{i,s}$ (a total of SN parameters). The choice of prior for the U_{std} represents the prior belief in the difference in judgment between the subjects. It can be set to be a weakly informative prior (with a large value for σ_U^2). The prior distribution for U_{std} is a half-normal distribution, i.e., normal distribution where only values above zero are valid.

Subject-specific predictors

The last extension presented in this article is the inclusion of subject-specific covariates. In many behavior research problems, it is desired to evaluate how characteristics of the subject influence the choice in a contest. This extension was originally proposed by Böckenholt (2001) models subject-specific covariates for each player utilizing a linear

regression. This model utilizes the following notation: K is the number of subject-specific predictors, $x_{i,k,s}$ representing the observed covariate k of subject s for player i and the coefficient for the covariate k of player i represented by $S_{i,k}$. The model can be represented as:

Likelihood:

$$\mathcal{P}[i \text{ beats } j] = \frac{\exp \lambda_{i,s}}{\exp \lambda_{i,s} + \exp \lambda_{j,s}} \tag{27}$$

$$\lambda_{i,s} = \lambda_i + \sum_{k=1}^{K} x_{i,k,s} S_{i,k} \tag{28}$$

$$y_{i,j} \sim \text{Bernoulli}(\mathcal{P}[i \text{ beats } j]) \tag{29}$$

Priors:

$$\lambda_i \sim \mathcal{N}(0, \sigma_\lambda^2) \tag{30}$$

$$S_{i,k} \sim \mathcal{N}(0, \sigma_S^2) \tag{31}$$

These models estimate the baseline ability of the players, λ_i, and the subject-specific coefficients of the covariates $S_{i,k}$. These coefficients represent how a change in the subject covariate for each player will impact the probability of selecting player i over player j. The model estimates one coefficient for every covariate of every player, resulting in a total of $K \cdot N$ estimated coefficients. It is worth noting that covariates coefficients are specific to each player and therefore can have a different impact, depending on how it influences the player log ability. These coefficients can be used to investigate systematic differences in how each player is evaluated by the subjects.

It is worth noting that again the absolute values of these coefficients do not have a direct interpretation of the effect it adds to the probability of one player beating another. This effect depends on the relative impact of the covariate in the two players and it is better assessed through the actual probability of selecting one player over the other. In the Bayesian context, this can be assessed through the posterior distribution of the probabilities and the absolute effect can be measured.

These models assume that the covariates $x_{i,k,s}$ have a similar range of values and that are centered, since they utilize the same normal priors with zero mean and constant standard deviation. In practice, this means that the values of the coefficients are more easily estimated by the MCMC sampler if all values of $x_{i,k,s}$ are normalized by each covariate. This model accepts both categorical predictors as well as continuous predictors. Categorical predictors can be added utilizing dummy-coding.

Remarks

Although not presented here, all the discussed extensions can be incorporated in a single-model mathematical model, since they are all linearly added in the exponential terms. The `bpcs` package can handle from both the simple Bradley–Terry and the Davidson model to any combination of these extensions to these models.

Even in more complex models the interpretation of the extension parameters remains the same as presented here. However, it is worth reinforcing that these parameters should always be analyzed in the context of the effect sizes, i.e., the actual probabilities of one player beating the other given the changes in the other parameters.

The bpcs package

This section presents a short overview of the underlying implementation of the `bpcs` package and its main functionalities. The `bpcs` R package implements the Bayesian version of the Bradley–Terry model and its extensions, as discussed in the statistical models' section. The models are coded in the Stan language and utilize the No-U-Turn (NUTS) Hamiltonian Monte Carlo sampler (Hoffman & Gelman, 2014), which provides several advantages over the Gibbs sampler (Nishio & Arakawa, 2019; Carpenter et al., 2017; Hoffman & Gelman, 2014). The latest version of the package and installation instructions can be found in the package repository.[2]

Basic usage

To exemplify the basic usage of the `bpcs` package, the work of Luckett, Burns, and Jenkinson (2020) is used as an example. The authors investigate the relative acceptability of food and beverage choices using paired preferences. One of the examples discussed is the acceptability of five brands of four frozen cheese pizzas. The full code for this presentation of the package and the reanalyses are available in the online appendix.

The main function of the `bpcs` package is the `bpc` function. The `bpc` function takes as input arguments: a data frame, two string columns with the names of contestants, a string with the result of the contestant (0 for player0, 1 for player1, or 2 for ties), and the model type. The model type is specified with a string. Two basic models are available, the '*bt*' model for the Bradley–Terry model (Bradley & Terry, 1952), and the '*davidson*' model for the Davidson model to handle ties (Davidson, 1970). Extensions for each of these base models can be added using a dash separator and the extension, for example, '*bt-ordereffect*' specifies the Bradley–Terry model with order effect; '*davidson-generalized-U*' specifies the generalized Davidson model including random effects. All presented extensions in the

[2] https://github.com/davidissamattos/bpcs

```
m <- bpc(data = dpizza,
         player0 = 'Prod0',
         player1 = 'Prod1',
         result_column = 'y',
         solve_ties = 'none',
         model_type = 'bt',
         iter=3000)
```

Listing 1 The bpc function

statistical models' section can be added to both base models, including more than one extension at the same time.

Other options, such as the method for handling ties, calculating the results from the scores of each player, column for clusters, specification of the priors, number of iterations to sample, among others, are described in the documentation.[3] The call for the `bpc` function is shown in the Listing 1:

The package also implements the S3 functions `print`, `summary`, `plot` and `predict`. The `print` function displays the parameters table with the High Posterior Density Intervals (Kruschke & Liddell, 2018; McElreath, 2020). `summary` function prints the parameters table, a table with a posterior probability of winning for all combination of players and a posterior rank of the players including the median rank, mean rank, and the standard deviation of the rank. The `plot` function provides a caterpillar plot of the model parameters with the correspondent HPD or credible intervals. The `predict` function provides a posterior distribution of predictive results of any match between the players of the fitted model. Below is the result of the `summary` function for the model.

The package also provides helper functions to create plots and to generate formatted tables (such as the ones from the `summary` function) for Latex, HTML, and the Pandoc[4] format (which in turn can be used to generate Microsoft Word tables). These functions are:

- `get_parameters_table` This function generates a table of the parameters with summary statistics of the posterior distribution. Two measures of uncertainty are available, the equal-tailed intervals and highest posterior density (HPD) intervals (Kruschke & Liddell, 2018; McElreath, 2020). The equal-tailed intervals divide the posterior distribution into two parts with the same probability mass, i.e., both tails have the same probability of being selected. The HPD interval corresponds to the narrowest interval that contains the mode for a unimodal distribution. In the case of a symmetrical unimodal distribution (such as the normal distribution), both intervals are equivalent. However, in the case of a non-symmetrical distribution, these intervals will be different and the HPD interval will be shorter.
- `get_probabilities_table`. This function generates a table of the probabilities of one player being chosen against another player. These probabilities are calculated by sampling the predictive posterior distribution of the results.
- `get_rank_of_players_table`. This function calculates the rank of each player based on the posterior distribution of the log abilities of the players (the λ). By assessing the posterior distribution of the rank and looking at the standard deviation, it is possible to assess the uncertainty on the rank estimates. Estimating the uncertainty in the rank values is not available in any of the frequentist packages.
- `plot`. This function creates a caterpillar-type of plot of the log-ability parameters of the players with the uncertainty intervals. This function returns a `ggplot2` (Wickham, 2016) object, which can be easily customized by the user.

If the user has a higher need to customize the tables, the user can either provide further customization with additional packages such as the `kableExtra` package (Zhu, 2020), or the utilize the function `get_parameters_df`, `get_probabilities_df` or `get_rank_of_players_df` to obtain a data frame that contains only the data of the table. The online appendix utilizes these approaches to create more complex tables for the reanalyses.

Model validity

After the call of the `bpc` function, the `bpcs` package runs the No-U-Turn Hamiltonian Monte Carlo sampler (Hoffman & Gelman, 2014) from Stan to estimate the posterior distribution of the parameters of the model. Before interpretation of the results, the user should check if the model has converged and or if there were problems in the convergence. If the model has not converged properly, the posterior distribution should not be interpreted. The basic checks are:

- Properly mixed chains. When sampling, it is common to use multiple chains. The chains should converge to the same value for every parameter and should not show any visible pattern (McElreath, 2020). A good convergence has a stationary caterpillar format. This can be checked using traceplots. Chains that have not converged in the presented paired comparison models are usually due to very large variance on the priors (which can lead to unidentifiable models since the soft-constraint is not sufficient).

[3] https://davidissamattos.github.io/bpcs/

[4] https://pandoc.org/

```
Estimated baseline parameters with 95

Table: Parameters estimates

Parameter              Mean    Median   HPD_lower   HPD_higher
-------------------  ------  -------  ----------  -----------
lambda[Tombstone]     -0.14    -0.11       -2.77         2.57
lambda[DiGiorno]       0.33     0.34       -2.30         3.03
lambda[Freschetta]     0.22     0.25       -2.42         2.97
lambda[Red Barron]     0.24     0.26       -2.41         2.95
lambda[aKroger]       -0.35    -0.32       -2.97         2.39
NOTES:
* A higher lambda indicates a higher team ability

Posterior probabilities:
These probabilities are calculated from the predictive posterior distribution
for all player combinations


Table: Estimated posterior probabilites

i            j             i_beats_j   j_beats_i
-----------  -----------  ----------  ----------
aKroger      DiGiorno           0.39        0.61
aKroger      Freschetta         0.35        0.65
aKroger      Red Barron         0.31        0.69
aKroger      Tombstone          0.47        0.53
DiGiorno     Freschetta         0.54        0.46
DiGiorno     Red Barron         0.48        0.52
DiGiorno     Tombstone          0.64        0.36
Freschetta   Red Barron         0.44        0.56
Freschetta   Tombstone          0.67        0.33
Red Barron   Tombstone          0.60        0.40

Rank of the players' abilities:
The rank is based on the posterior rank distribution of the lambda parameter

Table: Estimated posterior ranks

Parameter     MedianRank   MeanRank   StdRank
-----------  -----------  ---------  --------
DiGiorno               1       1.65      0.78
Freschetta             2       2.24      0.85
Red Barron             2       2.20      0.84
Tombstone              4       4.07      0.52
aKroger                5       4.84      0.37
```

Listing 2 Output of the summary function

- Gelman–Rubin convergence coefficient (split $\hat{R}$). This coefficient is another measure of convergence (Gelman & Rubin, 1992). A value close to 1 indicates indicate that the chains have converged to the same values. In practical terms, values of $\hat{R} < 1.01$ are required to indicate a good convergence McElreath (2020), and Vehtari, Gelman, Simpson, Carpenter, and Bürkner (2021) .
- Number of diverging iterations. Diverging iterations indicate that the sampler has not completely explored the solution space for the posterior. If there are diverging iterations, the results can be biased (Betancourt, 2017). A common solution is to increase the number of iterations for the warmup and the target acceptance probability parameter.
- Number of effective samples. This diagnoses the precision of the sampler estimation (Zitzmann & Hecht, 2019). The number of effective samples of the posterior indicates the number of independent samples. As a rule of thumb, 200 effective samples of the posterior are enough to estimate the mean of a parameter but more is required for estimating extreme quantiles (Zitzmann & Hecht, 2019; McElreath, 2020).

While these are the basic checks for any Monte Carlo sampler, there are two additional diagnostics specific to

Listing 3 Output of the check_convergence_diagnostics function

```
> check_convergence_diagnostics(m)

Checking sampler transitions treedepth.
Treedepth satisfactory for all transitions.

Checking sampler transitions for divergences.
No divergent transitions found.

Checking E-BFMI - sampler transitions HMC potential energy.
E-BFMI satisfactory for all transitions.

Effective sample size satisfactory.

Split R-hat values satisfactory all parameters.

Processing complete, no problems detected.
```

the Hamiltonian Monte Carlo (HMC) sampler used by the bpcs package.

- Maximum treedepth limits: the HMC imposes a limit in the depth of the trees that it evaluates at each iteration. If this limit is hit, it indicates that the sampler terminated to avoid long execution times. While it does not present a validity concern, the maximum treedepth represents an efficiency concern in terms of execution time. In the absence of other problems, increasing the treedepth may correct the problem (Stan Development Team, 2016).
- Low potential energy: a low kinetic energy calculated by the Estimated Bayesian Fraction of Missing Information (E-BFMI) indicates that the chains have not explored the posterior distribution efficiently. If this occurs, a common solution is to run the model for more iterations (Stan Development Team, 2016; Betancourt, 2017).

```
> get_waic(m)

Computed from 12000 by 380 log-likelihood matrix

          Estimate  SE
elpd_waic   -259.8 3.9
p_waic         4.0 0.1
waic         519.6 7.8
> get_loo(m)

Computed from 12000 by 380 log-likelihood matrix

         Estimate  SE
elpd_loo   -259.8 3.9
p_loo         4.0 0.1
looic       519.6 7.8
------
Monte Carlo SE of elpd_loo is 0.0.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

Listing 4 WAIC and LOO-CV in the bpcs package

The bpcs package offers basic convergence checks with the function check_convergence_diagnostics, as shown below. Figure 1 shows the traceplots for the pizza model.

These checks among other plots can also be verified with the shinystan package (Gabry, 2018). This package provides a web-based graphical user interface that implements convergence and posterior checks. The interface can be launched directly from the bpcs package with the function launch_shinystan.

Model comparison

Bayesian statistics reinforces generating several valid models that can explain the obtained data and comparing them (McElreath, 2020). One approach to comparing these models is with the use of an information criterion, such as the Watanabe-Akaike Information Criterion (WAIC) (Gelman, Hwang, & Vehtari, 2014) or the Leave-One-Out Cross-Validation method (LOO-CV) (Yao, Vehtari, Simpson, & Gelman, 2017).

The bpcs package provides both of these estimates with the functions get_waic and get_loo as shown below.

Note that the information criteria Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) should not be used since they assume models with flat priors and maximum a posteriori estimates (McElreath, 2020). These assumptions are not valid in the models implemented in the bpcs package and therefore the package does not provide these estimates.

Limitations of the bpcs package

The main limitation of these Bayesian paired comparison methods is the computational costs. While for most research problems a standard laptop should be able to create a posterior distribution of the parameters of the model in a few minutes, there are specific problems that will require

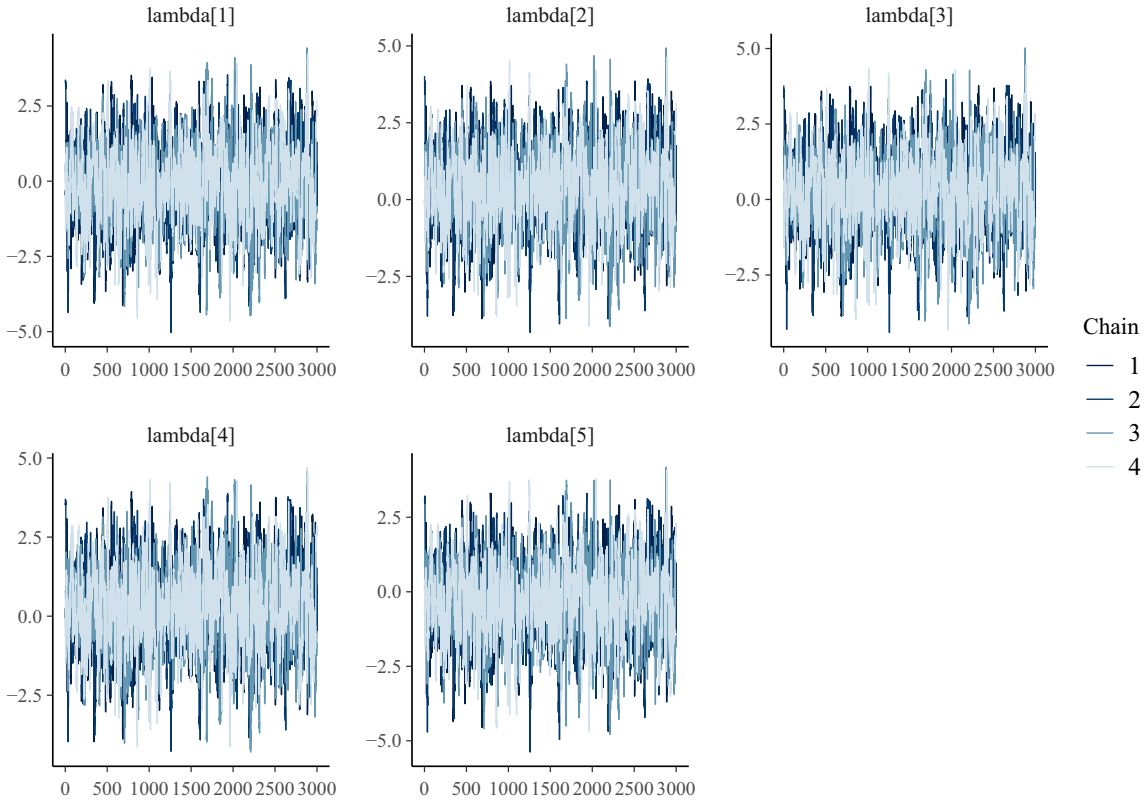

Fig. 1 Example of traceplots for the parameters of the pizza model. Note that the traceplots do not contain any apparent pattern (are stationary) and all chains are overlapping in a caterpillar format

more computational power, or even a transition from the Bayesian models towards frequentist models, if the posterior distribution is not used. For example, Zhang, Houpt, and Harel (2019) utilizes four Bradley–Terry models to rank a total of 7035 images with an order of 120,000 data points for each model. While it is still possible to perform Bayesian inference in this problem, fitting the model might take several hours. However, for many practical research problems, Bayesian Bradley–Terry models might take only minutes.

Reanalyses

This section provides Bayesian reanalyses of three studies conducted with a frequentist implementation of the Bradley–Terry model. The commented code to generate the figures and tables of these reanalyses are available in the Appendix. These reanalyses provide information regarding what was presented in the original papers, followed by discussions of alternative models. However, the reanalyses do not cover all possible models available in the `bpcs` package, such as the models with ties and generalized models. Examples of these models in other areas are provided in the package documentation[5].

Study I: Visual perception of moisture is a pathogen detection mechanism of the behavioral immune system

This reanalysis is based on the study "Visual Perception of Moisture Is a Pathogen Detection Mechanism of the Behavioral Immune System" (Iwasa, Komatsu, Kitamura, & Sakamoto, 2020). In this study, the authors utilized paired comparisons to rate the perceived moisture content based on the visual perception for high-luminance areas in images. The participants were asked to select the image that had the highest moisture content. The paired images were presented twice for each participant, first left to right and then right to left, to control for the influence of the presentation position. The image stimuli are presented in the supplementary material of the original article.

This reanalysis replicates the results of the study with a Bayesian Bradley–Terry model and then investigates the

[5] https://davidissamattos.github.io/bpcs/

Table 1 Parameters estimates for the simple Bradley–Terry model

Parameter	Mean	Median	HPD lower	HPD upper	N. Eff. Samples
image1	−4.577	−4.554	−6.533	−2.489	598
image2	−2.465	−2.436	−4.503	−0.461	596
image3	−0.154	−0.120	−2.221	1.830	596
image4	0.038	0.069	−2.065	1.977	596
image5	0.197	0.227	−1.874	2.175	595
image6	1.742	1.770	−0.349	3.668	594
image7	1.917	1.947	−0.188	3.842	594
image8	2.930	2.967	0.879	4.920	597

presence and magnitude of the order effect. The two models are compared utilizing the Watanabe–Akaike information criterion (WAIC).

For the simple Bradley–Terry model, Table 1 shows the worth value of each parameter that indicates the moisture content and the number of effective samples. Figure 2 shows the parameter plot with the 95% HPD intervals. The WAIC of the model is equal to 8132.2.

The second model adds an order-effect term to the model. The posterior estimation of the γ parameter is close to zero (with mean -0.001, lower HPD -0.054, and upper HPD 0.056), which indicates that there is no presence of order effect. The WAIC of the model is equal to 8134.1. The WAIC of the order effect is higher than the WAIC of the simple Bradley–Terry model, which indicates that the additional parameter did not increase the predictive values of the model. This indicates that, for this study, the strategy to show both images twice with change in the presentation order was effective to control for order effects. For the remaining analysis, the selected model is the Bradley–Terry model without order effect.

The priors were chosen to be normal distributions centered around 0 and with variance of 3.0. This variance allows probabilities of i beating j up to (given that each

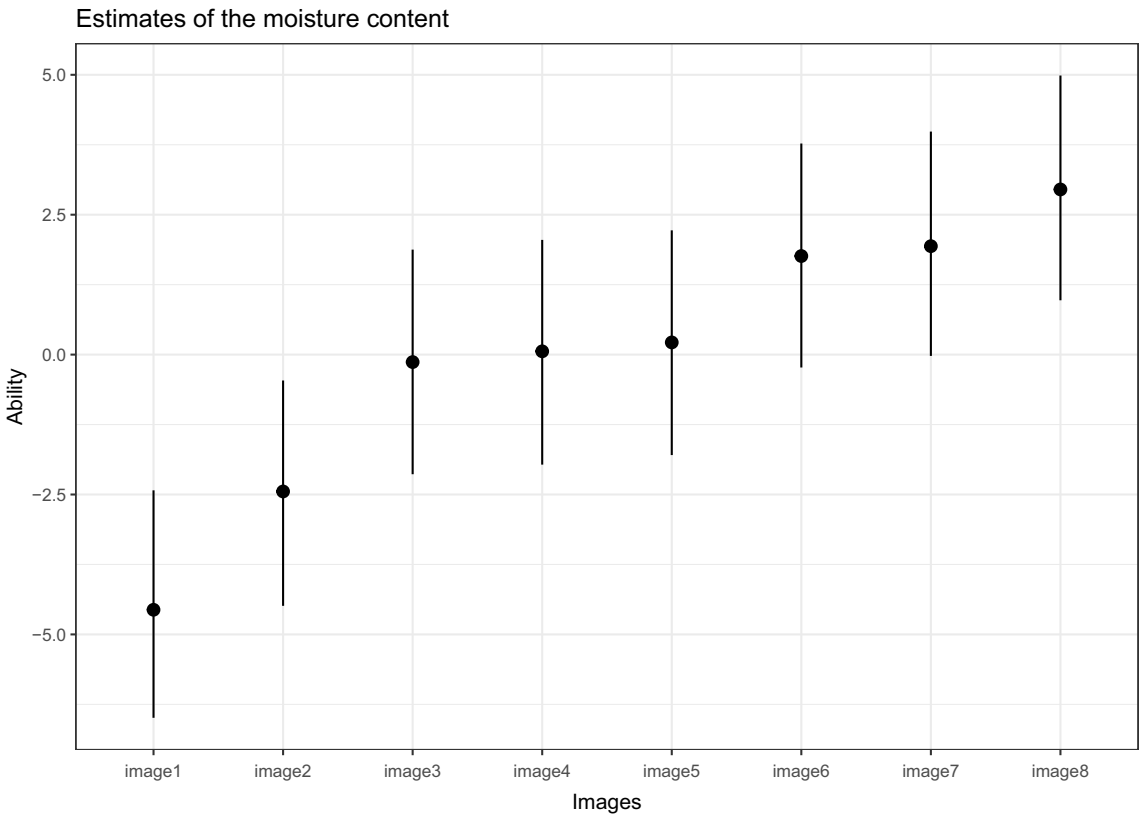

Fig. 2 Worth values of the images, in terms of moisture content, and their respective 95% HPD interval in the simple Bradley–Terry model

player can be up three standard deviations from the mean in each extreme) 0.99997. While this prior still regularizes and makes the model identifiable, it is considered weakly informative. The prior for the order effect γ parameter was set to mean zero and variance of 1.0. This prior indicates that the order effect can be for both the right or left images.

Considering the estimates of the first model in Fig. 2, it is possible to identify a large overlap in the interval between the latent worth value of some image groups (image3, image4, image5 and also image6 and image7). However, to properly rank and differentiate them, it is necessary to generate a posterior distribution of the rank of these images. Table 2 ranks the images based on the posterior distribution of the ranks in terms of moisture content. From this table, it is possible to see that despite the large overlap in the HPDI of the worth values, the images differentiate themselves in distinct ranks with low variation in the ranks.

Study II: Using a touchscreen paradigm to evaluate food preferences and response to novel photographic stimuli of food in three primate species

This reanalysis is from the article "Using a Touchscreen Paradigm to Evaluate Food Preferences and Response to Novel Photographic Stimuli of Food in Three Primate Species (*Gorilla gorilla gorilla*, *Pan troglodytes*, and *Macaca fuscata*)" (Huskisson, Jacobson, Egelkamp, Ross, & Hopper, 2020), an extension of the initial study with a single gorilla (Hopper, Egelkamp, Fidino, & Ross, 2019). In this study, the authors tested a protocol of pairwise forced choice with six stimuli of food (four familiar and two novel stimuli) for 18 subjects (six gorillas, five chimpanzees, and seven Japanese macaques). The study evaluates the efficacy of using touchscreens to test zoo-housed primates' food preferences and evaluate the understanding of the photographic stimuli.

A frequentist Bradley–Terry model was used to analyze food preference. The model was fitted with the `prefmod` (Hatzinger & Dittrich, 2012) and the `gnm` package (Turner & Firth, 2020) in R. The output of the analyses is the worth value of parameters. The analyses investigated species and subjects separately.

This reanalysis investigates a simple Bayesian Bradley–Terry model for each species without considering the multiple judgment sampling. Then, a simple model with random effects to model subject-specific preferences is performed. A total of six models were fitted (three simple Bradley–Terry and three Bradley–Terry with random effects). Table 3 shows the WAIC values for each model. This table indicates that all models with random effects perform better than the models without random effects since they have lower WAICs.

Similar to the previous reanalysis, the priors were chosen to be normal distributions centered around 0 and with variance of 3.0. While this prior still regularizes and makes the model identifiable, it is considered weakly informative. For the random effects, the variance was also set to 3.0, which allows large variances within the same cluster (in this example the individual primate) and being weakly informative.

Table 4 shows the obtained parameters for the random effects models together with HPD intervals. To complement, Fig. 3 shows a comparison of the estimates from the model with and without the random effects. Both models show a relatively close estimated value of the abilities of each food. Without considering the random effects term, the parameter value is equivalent to analyzing an average value for all individuals in the same cluster.

However, the effect sizes represented by the actual probability of choosing one food over the other for the species and the subjects can still be different. For example, the model with random effects can estimate the probability of each subject selecting one food over the other, while the Bradley–Terry model without random effects only estimates the species average. Additionally, the random-effects model can also compensate for non-balanced data, if one subject or species has more trials than another.

Two techniques can be used to assess the food preference. The first is through the posterior distribution of ranks of the foods. The second is through sampling the posterior distribution and calculating the probability of one stimulus

Table 2 Rank of the images based on moisture content

Parameter	Median Rank	Mean Rank	Std.Rank
image8	1	1.00	0.00
image7	2	2.00	0.05
image6	3	3.00	0.05
image5	4	4.01	0.08
image4	5	5.00	0.08
image3	6	6.00	0.03
image2	7	7.00	0.00
image1	8	8.00	0.00

Table 3 Comparison of the WAIC of the Bradley–Terry model and the Bradley–Terry model with random effects on the subjects for each species

	WAIC	
Specie	Bradley–Terry	Bradley–Terry with random effects
Macaques	7101.5	6712.6
Chimpanzees	7199.4	6720.0
Gorillas	9767.2	8792.9

Table 4 Parameters of the random effects model with 95% HPD and the number of effective samples

Parameter	Mean	Median	HPD lower	HPD upper	N. Eff. Samples
Macaque					
Carrot	0.12	0.12	−0.74	1.04	9934
Celery	−2.28	−2.29	−3.17	−1.39	9712
Jungle Pellet	1.23	1.23	0.36	2.15	9778
Oats	−0.90	−0.90	−1.77	0.05	9886
Peanuts	2.01	2.00	1.14	2.92	10195
Green Beans	−0.17	−0.17	−1.05	0.76	10176
Chimpanzees					
U1_std	0.58	0.57	0.42	0.75	4277
Apple	0.03	0.03	−0.97	1.08	10200
Tomato	0.32	0.32	−0.70	1.34	9880
Carrot	−0.21	−0.21	−1.25	0.77	9435
Grape	0.62	0.62	−0.38	1.69	10186
Cucumber	−0.32	−0.33	−1.36	0.67	9911
Turnip	−0.43	−0.43	−1.43	0.59	9342
Gorilla					
U1_std	0.72	0.70	0.49	1.01	4675
Apple	0.03	0.03	−0.95	0.99	13186
Carrot	−0.11	−0.11	−1.10	0.84	12365
Grape	0.86	0.87	−0.13	1.81	13237
Tomato	0.85	0.86	−0.12	1.85	13031
Cucumber	−0.70	−0.70	−1.67	0.27	12674
Turnip	−0.94	−0.94	−1.93	0.04	12944
U1_std	0.78	0.77	0.57	1.02	4883

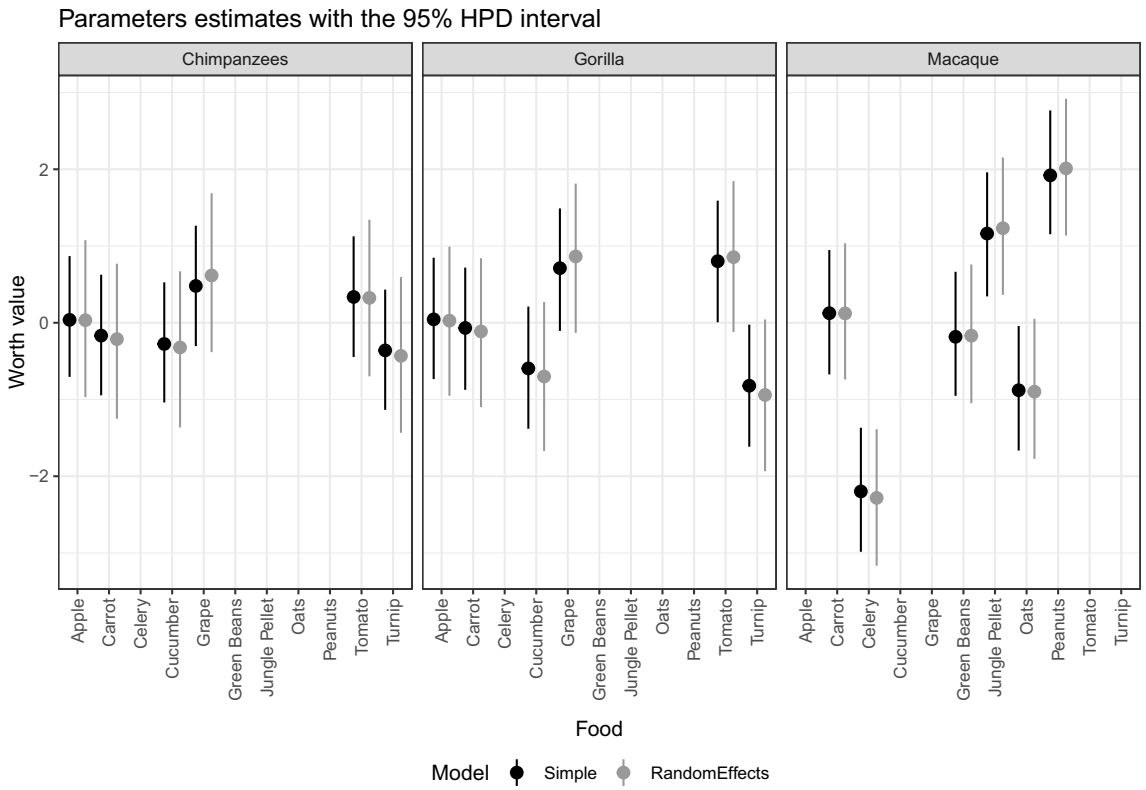

Fig. 3 The estimated abilities of each food type for each species in both models. Food items that do not have an estimated ability were not fed to that particular species

being chosen when compared to another. The second method represents a measure to assess the effect size of two competing stimuli.

Table 5 shows the rank of the food preferences for each species, the median, mean, and standard deviations. This rank is calculated from the posterior distribution of the ability parameters. This table indicates that the macaques have a well-defined rank for peanuts, jungle pellets, oats, and celery (given the low standard deviation of the rank). However, they do not have strong preferences between carrots and green beans. Chimpanzees have less consistent ranks and with higher standard deviation. They have a higher preference for grapes and a lower preference for turnip. Gorillas show a stronger preference for both grapes and tomatoes and a lower preference for turnips. The standard deviations on the ranks are smaller than with the chimpanzees that had the same food choice. It is worth noting, that this analysis consists of observing the preferences at the species level, not at the individual level. For example, although chimpanzees (at the species level) did not have well-defined ranks, each subject can have defined preferences, if analyzed individually.

The second method to assess food preference is with the posterior probability of selecting a stimulus over the other. The `bpcs` package provides functions to calculate these probabilities for all combinations except in the case of the random effect (in which the number of combinations is much larger). However, the package also offers the possibility to calculate the probability for selected cases.

In the case of the random-effects model, it is necessary to calculate the posterior distribution of each desired pair of stimuli for each subject. However, the package also has the capability to calculate the probabilities for the average of the subjects (the case in which random effects have a null effect in the probability). Table 6 shows the probabilities of selecting novel stimuli against the selection of old stimuli.

Table 5 Ranking of the food preferences per species for the random effects model

Food	Median Rank	Mean Rank	Std. Rank
Macaque			
Peanuts	1	1.01	0.09
Jungle Pellet	2	1.99	0.09
Carrot	3	3.17	0.37
Green Beans	4	3.84	0.39
Oats	5	4.99	0.08
Celery	6	6.00	0.00
Chimpanzees			
Grape	1	1.50	0.86
Tomato	2	2.24	1.10
Apple	3	3.36	1.25
Carrot	4	4.26	1.25
Cucumber	5	4.62	1.24
Turnip	5	5.01	1.10
Gorilla			
Tomato	1	1.54	0.60
Grape	2	1.57	0.61
Apple	3	3.37	0.68
Carrot	4	3.69	0.70
Cucumber	5	5.19	0.65
Turnip	6	5.65	0.57

Study III: Patients' health locus of control and preferences about the role that they want to play in the medical decision-making process

This reanalysis is from the article "Patients' health locus of control and preferences about the role that they want to play in the medical decision-making process" (Marton

Table 6 Posterior probabilities of the novel stimuli i being selected over the trained stimuli j

Item i	Item j	Probability	Odds Ratio
Gorilla			
Apple	Cucumber	0.61	1.56
Apple	Grape	0.23	0.30
Apple	Turnip	0.67	2.03
Apple	Carrot	0.56	1.27
Tomato	Cucumber	0.74	2.85
Tomato	Grape	0.50	1.00
Tomato	Turnip	0.87	6.69
Tomato	Carrot	0.75	3.00
Chimpanzee			
Apple	Cucumber	0.53	1.13
Apple	Grape	0.43	0.75
Apple	Turnip	0.54	1.17
Apple	Carrot	0.52	1.08
Tomato	Cucumber	0.64	1.78
Tomato	Grape	0.41	0.69
Tomato	Turnip	0.64	1.78
Tomato	Carrot	0.65	1.86
Macaque			
Oats	Celery	0.79	3.76
Oats	Jungle Pellet	0.16	0.19
Oats	Peanuts	0.05	0.05
Oats	Carrot	0.25	0.33
Green Beans	Celery	0.89	8.09
Green Beans	Jungle Pellet	0.19	0.23
Green Beans	Peanuts	0.09	0.10
Green Beans	Carrot	0.42	0.72

Table 7 Lambda parameters of the model and the random effects standard deviation

Parameter	Mean	Median	HPD lower	HPD lower	N. Eff. Samples
Active	−3.16	−3.14	−5.98	−0.52	2793
Active-Collaborative	2.10	2.08	−0.60	4.82	2764
Collaborative	4.88	4.89	2.01	7.71	2741
Passive-Collaborative	1.23	1.24	−1.38	4.00	2724
Passive	−5.11	−5.09	−7.99	−2.13	2718
U1_std	3.60	3.57	2.66	4.56	1899

et al., 2020). In the paper, the authors investigated how Health Locus of Control (HLOC) may influence patient's preferences for active or passive roles regarding their medical decision-making.

The study was conducted with 153 participants which responded to the Multidimensional Health Locus of Control Scale - form C (Ross, Ross, Short, & Cataldo, 2015) and a series of ten paired comparisons questions. The HLOC measured four dimensions (internal, chance, doctor, and other people) using an 18-point scale. The paired comparison questions were based on hypothetical situations of the Control Preference Scale - CPS (Solari et al. 2013) in which the participants chose one scenario among a set of comparative scenarios from Active role ("I prefer to make the decision about which treatment I will receive"); Active-Collaborative role ("I prefer to make the final decision about my treatment after seriously considering my doctor's opinion"); Collaborative role ("I prefer that my doctor and I share responsibility for deciding which treatment is best for me"); Passive-Collaborative ("I prefer that my doctor makes

Table 8 Subject predictors parameters by role

Parameter	Mean	Median	HPD lower	HPD lower	N. Eff. Samples
Active					
Internal	−0.16	−0.16	−2.72	2.61	2563
Chance	−0.15	−0.15	−2.93	2.56	2494
Doctors	−0.80	−0.80	−3.53	1.97	2211
Other people	−0.29	−0.28	−3.16	2.33	2576
Active-Collaborative					
Internal	−0.01	−0.01	−2.64	2.56	2534
Chance	−0.24	−0.25	−2.94	2.55	2374
Doctors	−0.74	−0.73	−3.50	1.92	2204
Other people	−0.50	−0.50	−3.39	2.14	2594
Collaborative					
Internal	−0.09	−0.08	−2.72	2.61	2571
Chance	0.09	0.09	−2.56	2.93	2362
Doctors	0.28	0.29	−2.48	3.02	2190
Other people	−1.22	−1.23	−3.98	1.51	2703
Passive-Collaborative					
Internal	0.12	0.12	−2.46	2.90	2516
Chance	0.13	0.14	−2.60	2.90	2355
Doctors	0.75	0.76	−1.97	3.54	2174
Other people	1.04	1.04	−1.69	3.81	2708
Passive					
Internal	0.00	0.01	−2.71	2.59	2509
Chance	−0.20	−0.20	−3.08	2.45	2375
Doctors	0.41	0.40	−2.38	3.12	2205
Other people	0.81	0.81	−1.90	3.59	2656

the final decision about which treatment will be used, but seriously considers my opinion"); or Passive role ("I prefer to leave all decisions regarding treatment to my doctor").

The data were analyzed with the frequentist Bradley–Terry model utilizing the `prefmod` package (Hatzinger & Dittrich, 2012). Four independent models were analyzed with each dimension of HLOC as the predictor, based on median-splitting to represent high HLOC and low HLOC for each dimension. The authors opted for this approach because the `prefmod` package only supports categorical predictors. This reanalysis evaluates three models in increasing complexity, with the four dimensions of HLOC modeled together. The models' fits are evaluated and how it impacts the estimated coefficients. The HLOC dimensions are normalized to both be presented at a comparable scale and to facilitate inference. Centering and scaling procedures such as normalizing facilitates the convergence of predictors coefficients McElreath (2020).

The first model is a simple Bradley–Terry model, to serve as a basis. This model has a WAIC of 1422.6. The second model utilizes the four dimensions of the HLOC as predictors and has a WAIC of 1378.7. The third model introduces both random effects to compensate for individual preferences for each of the five roles (active to passive) and the four HLOC dimensions as subject-specific predictors. This third model has a WAIC of 801.8 indicating the best fit out of the three models.

Similar to the previous reanalysis, the priors were chosen to be normal distributions centered around 0 and with variance of 3.0. While this prior still regularizes and makes

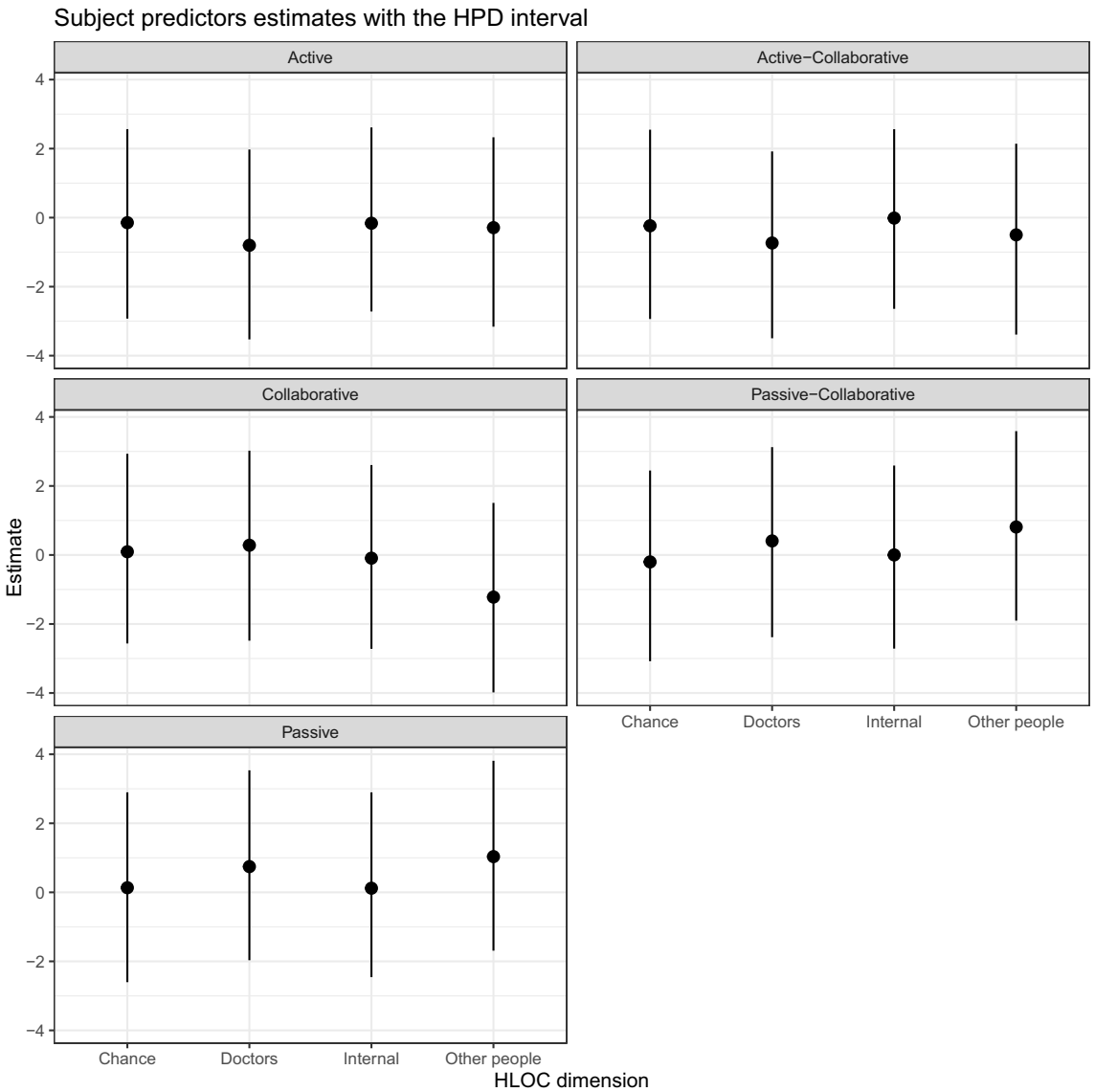

Fig. 4 The values of the subject predictors parameters with HPD intervals

the model identifiable it is considered weakly informative. For the random effects, the variance was also set to 3.0, which allows large variances within the same cluster and being weakly informative. For the subject specific predictors, the prior is also considered weakly informative and with a variance of 3.0.

Table 7 shows the values of the obtained λ parameters and the standard deviation of the random effects. The results indicated a higher base preference for the collaborative role and a lower base preference for the passive role.

Table 8 shows the parameters of the subject predictors. This table shows the values of the subject predictors parameters by each type of role. This table is more easily visualized with a plot, as shown in Fig. 4. Table 8 and Fig. 4 show the uncertainty in the actual impact of the HLOC in the CPS role. Most estimates have a median value close to zero and large HPD intervals overlapping zero. A similar conclusion can also be assessed in the probabilities of a selecting a specific CPS role, as shown in Table 9.

Table 9 shows a few cases illustrating the impact of the HLOC dimensions in the actual probability of selecting a specific CPS role. Specifically, this table shows the probabilities of a subject to select between the roles active and passive, active-collaborative and collaborative, and collaborative and passive-collaborative with changes of the HLOC in the different dimensions. This table shows that the probability choice between the roles active and passive for the mean value is 0.88. For a subject with internal HLOC two standard deviations below the average this probability changes to 0.80, when the internal HLOC is two standard deviations above this probability in unchanged compared to the average. This indicates that the internal HLOC has a small impact on the selection of these two roles. These small changes in probabilities indicate that while using HLOC dimensions as subject predictors improves the model, their effect on the actual probabilities of selecting a role are small. The baseline coefficients for each role contribute more relative difference between the roles than the effect of the subject-specific predictors.

Conclusion

The ultimate goal of this article is to provide tools with strong theoretical foundations that empower researchers to have alternatives to the use of frequentist data analysis when analyzing paired data. Therefore, the `bpcs` package was introduced to facilitate the adoption of Bayesian models in paired comparison assessments. The package is free to use, and the latest version is available at the official repository.

This article explained the rationales behind the different Bayesian models implemented in the `bpcs` package. Additionally, the article provides the reanalyses of three studies in various areas of behavior science (psychophysics, animal research, and health). The package allows researchers to run the Bayesian Bradley–Terry model and many of its extensions, such as the Davidson model to handle ties, models with order effect, generalized models, models with dependent data, and models with predictors on the subject (and the different combinations of these extensions). It also provides

Table 9 Probabilities of selecting role i instead of j based on changes of the values of the HLOC dimensions

Roles		HLOC dimensions				
i	j	Internal	Chance	Doctors	OtherPeople	Probability
Active	Passive	0	0	0	0	0.88
Active	Passive	−2	0	0	0	0.80
Active	Passive	2	0	0	0	0.88
Active	Passive	0	0	−2	0	0.89
Active	Passive	0	0	2	0	0.78
Act.-Collab.	Collab.	0	0	0	0	0.08
Act.-Collab.	Collab.	−2	0	0	0	0.08
Act.-Collab.	Collab.	2	0	0	0	0.07
Act.-Collab.	Collab.	0	0	−2	0	0.07
Act.-Collab.	Collab.	0	0	2	0	0.04
Collab.	Pass.-Collab.	0	0	0	0	0.98
Collab.	Pass.-Collab.	0	−2	0	0	0.94
Collab.	Pass.-Collab.	0	2	0	0	0.98
Collab.	Pass.-Collab.	0	0	0	−2	0.97
Collab.	Pass.-Collab.	0	0	0	2	0.96

tools for assessing uncertainty in the ranks and the posterior probabilities, not available in frequentist packages. All the code used to fit the models and create the tables and the figures from the reanalyses section are available in the online appendix.

Being able to easily extend a simple model to more complex ones, as shown in the reanalyses, allows researchers to control bias and errors in the modeling. Future research could further develop Bayesian cumulative models (when there is a strength scale in the assessment of two items) and models with time dependency.

Open practices statement

This study was not pre-registered. The data used in the reanalyses can be accessed in the referenced publication or upon request to the authors. The bpcs package is open-source under the MIT license and the source code can be accessed at the package repository: https://github.com/davidissamattos/bpcs.

Funding Open access funding provided by Chalmers University of Technology. This work was partially supported by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation.

The authors would like to thank L. Hopper for revising the results of the second reanalysis and G. Marton for revising the results of the third reanalysis.

References

Abalos, J., de Lanuza, G. P., Carazo, P., & Font, E. (2016). The role of male coloration in the outcome of staged contests in the European common wall lizard (*Podarcis muralis*). *Behaviour*, *153*(5), 607–631.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434.

Böckenholt, U. (2001). Hierarchical modeling of paired comparison data. *Psychological Methods*, *6*(1), 49.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*(3/4), 324–345.

Brown, A. (2016). Item response models for forced-choice questionnaires: a common framework. *Psychometrika*, *81*(1), 135–160.

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*, *71*(3), 460–502.

Bush, J. M., Quinn, M. M., Balreira, E. C., & Johnson, M. A. (2016). How do lizards determine dominance? Applying ranking algorithms to animal social behaviour. *Animal Behaviour*, *118*, 65–74.

Butler, K., & Whelan, J. T. (2004). The existence of maximum likelihood estimates in the Bradley–Terry model and its extensions. arXiv:math/0412232.

Caron, F., & Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, *21*(1), 174–196.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., & et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Cattelan, M. (2012). Models for paired comparison data: a review with emphasis on dependent data. *Statistical Science*, 412–433.

Chien, S. H.-L., Lin, Y.-L., Qian, W., Zhou, K., Lin, M.-K., & Hsu, H.-Y. (2012). With or without a hole: Young infants' sensitivity for topological versus geometric property. *Perception*, *41*(3), 305–318.

Coetzee, H., & Taylor, J. (1996). The use and adaptation of the paired-comparison method in the sensory evaluation of hamburger-type patties by illiterate/semi-literate consumers. *Food Quality and Preference*, *7*(2), 81–85.

Corff, S. L., Lerasle, M., & Vernet, E. (2018). A Bayesian nonparametric approach for generalized Bradley–Terry models in random environment. arXiv:1808.08104.

Davidson, R. R. (1970). On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, *65*(329), 317–328.

Davidson, R. R., & Beaver, R. J. (1977). On extending the Bradley–Terry model to incorporate within-pair order effects. *Biometrics*, 693–702.

Davidson, R. R., & Solomon, D. L. (1973). A Bayesian approach to paired comparison experimentation. *Biometrika*, *60*(3), 477–487.

Dittrich, R., Hatzinger, R., & Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(4), 511–525.

Fleischhaker, D. S. (2019). Modelling outcomes in Canadian professional football via generalized Bradley–Terry models. Unpublished doctoral dissertation, The University of Regina (Canada).

Ford, J.r., L. R. (1957). Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, *64*(8P2), 28–33.

Gabry, J. (2018). Shinystan: Interactive visual and numerical diagnostics and posterior analysis for Bayesian models [Computer software manual]. https://CRAN.R-project.org/package=shinystan (R package version 2.5.0).

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Giambona, F., & Grassini, L. (2020). Tourism attractiveness in Italy: Regional empirical evidence using a pairwise comparisons modelling approach. *International Journal of Tourism Research*, *22*(1), 26–41.

Glickman, M. E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, *28*(6), 673–689.

Hägerhäll, C. M., Ode Sang, Å., Englund, J.-E., Ahlner, F., Rybka, K., Huber, J., & et al. (2018). Humans really prefer semi-open natural landscapes? A cross-cultural reappraisal. *Frontiers in psychology*, *9*, 822.

Handley, J. C. (2001). Comparative analysis of Bradley–Terry and Thurstone–Mosteller paired comparison models for image quality assessment. In *Pics*, (Vol. 1, pp. 108–112).

Hatzinger, R., & Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, *48*(10), 1–31.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hontangas, P. M., Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). De la Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement*, *39*(8), 598–612.

Hopper, L. M., Egelkamp, C. L., Fidino, M., & Ross, S. R. (2019). An assessment of touchscreens for testing primate food preferences and valuations. *Behavior Research Methods*, *51*(2), 639–650.

Huskisson, S. M., Jacobson, S. L., Egelkamp, C. L., Ross, S. R., & Hopper, L. M. (2020). Using a touchscreen paradigm to evaluate food preferences and response to novel photographic stimuli of food in three primate species (*Gorilla gorilla gorilla*, *Pan troglodytes*, and *Macaca fuscata*). *International Journal of Primatology*, 1–19.

Iwasa, K., Komatsu, T., Kitamura, A., & Sakamoto, Y. (2020). Visual perception of moisture is a pathogen detection mechanism of the behavioral immune system. *Frontiers in Psychology*, *11*, 170.

Johnson, T. R., & Kuhn, K. M. (2013). Bayesian Thurstonian models for ranking data using JAGS. *Behavior Research Methods*, *45*(3), 857–872.

Kelter, R. (2020). Analysis of type I and II error rates of Bayesian and frequentist parametric and nonparametric two-sample hypothesis tests under preliminary assessment of normality. *Computational Statistics*, 1–26.

Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: forced-choice vs. psychometric modeling of Likert items. *Frontiers in Psychology*, *10*, 2309.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin and Review*, *25*(1), 155–177.

Kucukelbir, A., Ranganath, R., Gelman, A., & Blei, D. (2015). Automatic variational inference in Stan. *Advances in Neural Information Processing Systems*, *28*, 568–576.

Leonard, T. (1977). An alternative Bayesian approach to the Bradley–Terry model for paired comparisons. *Biometrics*, 121–132.

Luckett, C. R., Burns, S. L., & Jenkinson, L. (2020). Estimates of relative acceptability from paired preference tests. *Journal of Sensory Studies*, *35*(5), e12593.

Marton, G., Pizzoli, S. F. M., Vergani, L., Mazzocco, K., Monzani, D., Bailo, L., & et al. (2020). Patients' health locus of control and preferences about the role that they want to play in the medical decision-making process. *Psychology, Health and Medicine*, 1–7.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton: CRC Press.

Meid, A. D., Quinzler, R., Groll, A., Wild, B., Saum, K.-U., Schöttker, B., et al. (2016). Longitudinal evaluation of medication underuse in older outpatients and its association with quality of life. *European Journal of Clinical Pharmacology*, *72*(7), 877–885.

Miller, E. T., Bonter, D. N., Eldermire, C., Freeman, B. G., Greig, E. I., Harmon, L. J., et al. (2017). Fighting over food unites the birds of North America in a continental dominance hierarchy. *Behavioral Ecology*, *28*(6), 1454–1463.

Morillo, D., Leenen, I., Abad, F. J., Hontangas, P., De la Torre, J., & Ponsoda, V. (2016). A dominance variant under the multi-unidimensional pairwise-preference framework: Model formulation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *40*(7), 500–516.

Nishio, M., & Arakawa, A. (2019). Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution*, *51*(1), 1–12.

Petrou, S. (2003). Methodological issues raised by preference-based approaches to measuring the health status of children. *Health Economics*, *12*(8), 697–702.

Phelan, G. C., & Whelan, J. T. (2017). Hierarchical Bayesian Bradley–Terry for applications in major league baseball. arXiv:1712.05879.

Pritikin, J. N. (2020). An exploratory factor model for ordinal paired comparison indicators. *Heliyon*, *6*(9), e04821.

Ross, T. P., Ross, L. T., Short, S. D., & Cataldo, S. (2015). The multidimensional health locus of control scale: Psychometric properties and form equivalence. *Psychological reports*, *116*(3), 889–913.

Seymour, R. G., Sirl, D., Preston, S., Dryden, I. L., Ellis, M. J., Perrat, B., & et al. (2020). The Bayesian spatial Bradley–Terry model: urban deprivation modeling in Tanzania. arXiv:2010.14128.

Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K., & Wainwright, M. (2015). Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, (pp. 856–865).

Solari, A., Giordano, A., Kasper, J., Drulovic, J., van Nunen, A., Vahter, L., et al. (2013). Role preferences of people with multiple sclerosis: image-revised, computerized self-administered version of the control preference scale. *PLoS One*, *8*(6), e66127.

Sport (2020). Sport: an R package for online ranking methods. https://github.com/gogonzo/sport, (R package version 0.2.0.

Springall, A. (1973). Response surface fitting using a generalization of the Bradley–Terry paired comparison model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *22*(1), 59–68.

Stan Development Team (2016). Stan modeling language users guide and reference manual. Technical report.

Stern, S. E. (2011). Moderated paired comparisons: a generalized Bradley–Terry model for continuous data using a discontinuous penalized likelihood function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*(3), 397–415.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273.

Tsukida, K., & Gupta, M. R. (2011). How to analyze paired comparison data (Tech. Rep.). Washington Univ Seattle Dept Of Electrical Engineering.

Turner, H., & Firth, D. (2012). Bradley–Terry models in R: the BradleyTerry2 package. *Journal of Statistical Software, 48*(9).

Turner, H., & Firth, D. (2020). Generalized nonlinear models in R: An overview of the gnm package [Computer software manual]. https://cran.r-project.org/package=gnm, (R package version 1.1-1).

Turner, H. L., van Etten, J., Firth, D., & Kosmidis, I. (2020). Modelling rankings in R: the PlackettLuce package. *Computational Statistics*, 1–31.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved R for assessing convergence of MCMC. *Bayesian Analysis*, *1*(1), 1–28.

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement*, *41*(8), 600–613.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer. https://ggplot2.tidyverse.org.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2017). Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*.

Zhang, H., Houpt, J. W., & Harel, A. (2019). Establishing reference scales for scene naturalness and openness. *Behavior Research Methods*, *51*(3), 1179–1186.

Zhu, H. (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax [Computer software manual]. https://CRAN.R-project.org/package=kableExtra (R package version 1.2.1).

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 646–661.