Methods for Model Comparison in Cognitive Modeling (Postprint)

Authors: Guo Mingqian, Pan Wanke, Hu Chuanpeng, Guo Mingqian, Hu Chuanpeng

Date: 2024-10-30T00:00:00+00:00

Abstract

Cognitive modeling has gained widespread application in scientific psychology in recent years, and model comparison constitutes a critical component of cognitive modeling: researchers must select an optimal model through model comparison before proceeding with subsequent hypothesis testing or latent variable inference. Model comparison must consider not only the model's fit to the data (balancing overfitting and underfitting), but also the complexity of parameters and mathematical form. However, model comparison indices are numerous and complex. We divide the commonly used model comparison indices in cognitive modeling into three major categories, introducing their calculation methods, advantages, and disadvantages, including goodness-of-fit indices (including mean squared error, coefficient of determination, ROC curve, etc.), cross-validation-based indices (including AIC, DIC, etc.), and marginal likelihood-based indices. Combining simulated and real data from the orthogonal Go/No-Go paradigm, we demonstrate how each index can be implemented in R. On this basis, we discuss the applicable contexts of each index and introduce new approaches to model comparison such as model averaging.

Full Text

Methods for Model Comparison in Cognitive Modeling

Mingqian Guo[1], Wanke Pan[2], Chuan-Peng Hu[2]
([1] Behavioral Science Institute, Radboud University, Nijmegen 6525GD, The Netherlands;
[2] School of Psychology, Nanjing Normal University, Nanjing 21002, China)

Abstract: Cognitive modeling has gained widespread application in scientific psychology in recent years, and model comparison constitutes a critical component of this process: researchers must select an optimal model through model

comparison before proceeding with subsequent hypothesis testing or latent variable inference. Model comparison requires consideration not only of model fit to data (balancing overfitting and underfitting) but also of complexity arising from parameter data and mathematical form. However, numerous model comparison indices exist, creating considerable complexity. This article categorizes commonly used model comparison indices in cognitive modeling into three major classes, introducing their calculation methods, advantages, and disadvantages: goodness-of-fit metrics (including mean squared error, coefficient of determination, ROC curves, etc.), cross-validation-based metrics (including AIC, DIC, etc.), and marginal likelihood-based metrics. Using simulated and real data from the orthogonal Go/No-Go paradigm, we demonstrate how to implement each metric in R. Building on this foundation, we discuss the appropriate contexts for each metric and introduce novel approaches such as model averaging.

Keywords: Cognitive modeling; Computational models; Model comparison; Model selection

Over the past two decades, research employing computational models for cognitive modeling of behavioral data has attracted increasing attention from researchers. For example, in the domain of perceptual decision-making, Bayesian perception models (Kording & Wolpert, 2006) and drift diffusion models (Forstmann et al., 2016; Ratcliff et al., 2016) have been widely applied in cognitive neuroscience. Similarly, reinforcement learning models have become increasingly mainstream in value-based decision-making research, with model-estimated latent variables such as "prediction error" effectively predicting dopaminergic neuron activity during learning (Schultz et al., 1997; Steinberg et al., 2013). Computational models also form the foundation of the emerging interdisciplinary field of computational psychiatry (Geng et al., 2022; Huys et al., 2016; Montague et al., 2012; 区健新, 2020), enhancing understanding of cognitive processing deficits in psychiatric populations to improve diagnostic accuracy and classification, and enabling precision treatment (Pedersen et al., 2021).

The cognitive modeling process typically includes steps such as simulation, parameter estimation, model comparison, and latent variable inference (Wilson & Collins, 2019). Specifically, researchers propose computational models based on different theories, conduct simulations, design experiments to collect data, fit data with various computational models, select the optimal model through model comparison, and finally analyze the data based on the optimal model, combining the model's latent variables with neural data for inference.

Model comparison is a crucial step in cognitive modeling, essential not only within cognitive modeling but also in any scenario involving computational models. However, researchers in psychology and cognitive science are often unfamiliar with the model comparison process and frequently feel confused when faced with the wide variety of model comparison indices. Furthermore, current literature lacks systematic organization of the many methods for model comparison. In light of this, this article reviews the principles and methods of model

comparison to help readers understand the underlying rationale and appropriate contexts for current model comparison practices, thereby promoting better application of cognitive modeling. While this article focuses on cognitive modeling in experimental psychology, the introduced metrics can also be applied to other common statistical models in psychology, such as hierarchical linear regression and structural equation modeling.

We will first introduce the basic principles of model comparison, then systematically review the rationale, advantages, and disadvantages of common model comparison indices using case examples, and finally summarize the strengths, weaknesses, and usage considerations of each index from a practical application perspective.

1 Basic Principles of Model Comparison

For researchers, a good model must possess two key characteristics. First, it must adequately explain or fit the current sample data. Second, it must have generalization capability—that is, it must provide good explanations for data beyond the current sample (i.e., predictive ability). If a model cannot accurately explain the current sample data, it is considered underfitting. If a model explains the current sample data very well but fails to explain out-of-sample data, it is considered overfitting (Friedman et al., 2001).

Researchers typically use generalization error, which measures the difference between model predictions and real data, to assess a model's generalization capability. Generalization error can be decomposed into variance, bias, and irreducible error. Bias measures the deviation between the expected value of model predictions and the true data. A model with high bias typically means the model is too simple to capture complex relationships in the data, leading to underfitting. Variance measures the variability of prediction results across different training datasets. A model with high variance typically means the model is too complex and has learned random noise in the training data, leading to overfitting. Irreducible error refers to the unavoidable noise and uncertainty inherent in the data itself. This error arises from the complexity of the data itself or measurement errors, and no model can predict or eliminate this portion of error. Therefore, as shown in Figure 1 [Figure 1: see original paper], as model complexity increases, model bias gradually decreases while variance increases, a phenomenon known as the bias-variance trade-off. Models with high bias underfit, while models with high variance overfit (Friedman et al., 2001). Model selection is a process of trading off bias and variance to minimize generalization error.

Figure 1. Schematic diagram of the bias-variance trade-off. As model complexity increases, bias gradually decreases while variance gradually increases. Total error has a minimum value.

Although model complexity plays an important role in generalization capability, it is influenced by several factors. Myung and Pitt (1997) summarized

three factors affecting model complexity. The first is the number of model parameters—generally, more parameters mean higher complexity. The second is mathematical form—for example, nonlinear models are more complex than linear models. The third is the range of the parameter space—a larger parameter space range indicates more degrees of freedom and thus greater complexity.

Based on differences in focus and rationale, model comparison indices can be divided into three categories. The first category is goodness-of-fit, which does not consider model complexity but simply measures how well the model fits the current sample data. The second category includes cross-validation and approximate cross-validation indices, which focus on generalization ability—that is, out-of-sample prediction accuracy based on models fitted to current sample data. The third category is based on marginal likelihood, (|), where represents observed data and represents the model. Marginal likelihood focuses on selecting the "true model" that may exist among candidate models. The latter two categories both balance complexity and goodness-of-fit. Different model comparison indices have their own advantages and disadvantages, and no single index is universally superior. Therefore, researchers need to select appropriate indices based on actual circumstances. The following sections will use a dataset as an example to introduce these three major categories of indices.

Notably, the method used to fit cognitive models also influences the choice of model indices, as some indices are only applicable with specific fitting methods. Figure 2 [Figure 2: see original paper] shows the model comparison indices corresponding to different fitting methods. Methods for fitting cognitive models include point estimation methods such as Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP) estimation, as well as Bayesian parameter estimation, which estimates the entire posterior distribution rather than relying on point estimates. Bayesian parameter estimation offers distinct advantages. First, Bayesian estimation provides posterior distributions of parameters, which not only facilitates subsequent analysis but is particularly beneficial for building hierarchical models. The prior distribution in Bayesian parameter estimation serves a regularization function, thereby reducing model complexity (Bishop, 2006). Additionally, Bayesian methods demonstrate unique advantages when handling data from multiple subjects. Bayesian estimation is highly conducive to building hierarchical Bayesian models, which introduce group-level priors. Different subjects' parameters are drawn from a distribution formed by group-level parameters, while the estimation of group-level parameters itself is constrained by individual subject parameters. Consequently, individual subject parameter values are indirectly influenced by other subjects' data through group-level parameters, shifting toward the group-level parameter mean and thereby reducing the impact of extreme data from individual subjects (Ahn et al., 2017; Gelman, Carlin, et al., 2013).

Figure 2. Three common classes of model comparison indices in cognitive modeling, including goodness-of-fit metrics, cross-validation-based metrics, and marginal likelihood-based metrics.

2 Goodness-of-Fit Metrics

Goodness-of-fit primarily measures the degree of prediction or fit of a model to experimental data. Although goodness-of-fit indices do not account for overfitting caused by increasing model complexity, their role in cognitive modeling should not be overlooked. First, goodness-of-fit indices can be used to explore a model's absolute performance. Second, these indices can be used to compare models when complexity differences are small and when nested models exist. Commonly used goodness-of-fit indices in cognitive modeling include: Mean Squared Error (MSE), Coefficient of Determination (^{2}/pseudo- ^{2}), Log Likelihood Function, Receiver Operating Characteristic (ROC) curves, and Posterior Predictive Checks. Table 1 summarizes the advantages and disadvantages of each index.

Table 1. Advantages, disadvantages, and applicable parameter estimation ranges of various goodness-of-fit metrics

Metric	Applicable Parameter Estimation Methods	Advantages	Disadvantages
MSE	Maximum likelihood, least squares	Intuitive and simple, easy to calculate and understand	Not applicable to classification problems, does not consider model complexity, prone to overfitting
Coefficient of determination (^{2})	Maximum likelihood, least squares	Measures proportion of variability explained, provides interpretable fit measure	Sensitive to model complexity, cannot compare models with different numbers of features

Metric	Applicable Parameter Estimation Methods	Advantages	Disadvantages
Log likelihood function	Maximum likelihood, MAP, Bayesian estimation	Reflects match between model predictions and actual data; can be used for model comparison and parameter estimation; MSE and ^{2} are special cases of log likelihood under normal residuals	Not applicable to non-probabilistic, non-parametric models; sensitive to outliers
ROC curve	Maximum likelihood, MAP, Bayesian estimation	Used to evaluate model's ability to predict actual data	Not applicable to multi-option data; for imbalanced data, results are less accurate
Posterior predictive check	Bayesian parameter estimation	Considers parameter uncertainty and model complexity; can check prediction ability for new data samples	Requires domain expertise to specify prior and posterior distributions; high computational complexity

2.1 Mean Squared Error

Mean Squared Error, abbreviated as MSE (Mean Squared Error), also known as Mean Squared Deviation (MSD), is a common metric for evaluating general linear regression. Its calculation formula is:

$$MSE = \sum (yi - yi)^2$$

where is a data point from the sample and ˆ is the model's predicted value. MSE is typically applied to regression prediction problems where the modeled data are continuous variables. MSE is not suitable for classification problems like the case study in this article.

Taking the square root of MSE yields Root Mean Square Deviation (RMSD). Multiplying MSE by the number of data points yields Residual Sum of Squares (RSS). When models use a Gaussian distribution, RSS can be used for F-tests of nested models. Nested models refer to models that have fewer parameters or where certain parameters are constrained (e.g., fixed to specific values) relative to another model. In nested models, one model (the simpler model) is a subset of another model (the more complete model), reducing complexity based on the more complete model.

The F-statistic formula is:

$$\frac{RSS_{reduced} - RSS_{full}}{RSS_{full}} \cdot \frac{df_{full}}{\Delta p}$$

where $RSS_{reduced}$ and RSS_{full} are the RSS values for the simple and full models, respectively, Δp is the difference in free parameters between them, and df_{full} is the degrees of freedom for the full model (Hair et al., 2010). Additionally, RSS from a Gaussian distribution can substitute for the log likelihood function when calculating AIC and BIC (Friedman et al., 2001; Lebreton et al., 2019). More about AIC and BIC can be found in Sections 3.1 and 4.1, respectively.

2.2 Coefficient of Determination

The coefficient of determination ^{2} is commonly used to measure goodness-of-fit in linear regression models. The value of ^{2} ranges from 0 to 1, reflecting the proportion of variance in the dependent variable explained by independent variables. The closer ^{2} is to 1, the better the model fits the data. Its calculation formula is:

$$r^2 = 1 - \frac{RSS}{TSS}$$

where TSS (Total Sum of Squares) is the total sum of squares and RSS (Residual Sum of Squares) is the residual sum of squares. Their calculation formulas are:

$$TSS = \sum (yi - \bar{y})^2$$
$$RSS = \sum (yi - \hat{yi})^2$$

Like MSE, the coefficient of determination ^{2} is commonly applied to regression prediction problems where modeled variables are continuous and is not suitable for classification problems with discrete distributions like the case study in this article.

To make ^{2} applicable to discrete distributions, researchers have proposed using pseudo- ^{2}. There are multiple formulas for pseudo- ^{2}; this article introduces one proposed by McFadden (1984) as an example because it satisfies eight desirable properties of coefficients of determination identified by Kvålseth (1985) (Menard, 2000).

The formula is:

$$pseudo\ r^2_{McFadden} = 1 - \frac{\sum LL_{full\ model}}{\sum LL_{null\ model}}$$

where $\sum LL_{full\ model}$ is the sum of log likelihood functions for the model, and $\sum LL_{null\ model}$ is the sum for the null model (Daw, 2011; McFadden, 1984). The null model assumes that experimental stimuli have no effect on observed data and that observed data are uniformly distributed. Here, the null model refers to a binomial or multinomial distribution model with parameters equal to (1/number of options). For example, in the case study in this article, there are two possible options, so the binomial distribution parameter is 0.5, meaning equal probability of observing both options, and the null model's likelihood function is the number of trials multiplied by log(0.5).

2.3 Log Likelihood Function

The likelihood function represents the probability that each model parameter generates the observed data given the observed data. Taking the logarithm of the likelihood function yields the log likelihood function, which can be used to assess the fit between model parameters and actual data and is typically used in Maximum Likelihood Estimation (MLE). The likelihood function formula is:

$$\log L(\theta|y) = \log p(y|\theta)$$

Different tasks have different data distributions, so the form of the log likelihood function also varies. For choice data, the log likelihood function is typically constructed based on Bernoulli or multinomial distributions; for continuous data such as reaction times or EMG, it is generally constructed based on Gaussian distributions (Ballard et al., 2019; Ikink et al., 2019; Li et al., 2011).

In cognitive modeling model comparison, the log likelihood function typically serves two purposes. First, the average log likelihood function is used to investigate a model's absolute performance. The example in this article is a binary choice task, where the probability of random selection is 50% with a log value of -0.693. Therefore, when the average log likelihood function is greater than -0.693, the model's performance is better than chance level.

Second, the log likelihood can be used to calculate the Likelihood Ratio Test (LRT) to infer whether performance differences between nested models are significant. The asymptotic distribution of the LRT is chi-square, with degrees of freedom proportional to the difference in the number of free parameters between the two models (Casella & Berger, 2002; Wilks, 1938).

The LRT formula is:

$$LRT = -2 \times (\log L_{reduced} - \log L_{full})$$

where L_{full} is the likelihood function for the full model and $L_{reduced}$ is for the model with certain parameters fixed. In specific calculations, we need to sum the likelihood functions across all trials for all subjects to compute the LRT and check the chi-square distribution to determine whether model differences are significant. In the case study, we used the LRT to compare Model 1 and Model 2. The difference in free parameters between these two models is 2, multiplied by the number of subjects (61), so a chi-square distribution with 121 degrees of freedom can be used for the LRT. The p-value for the LRT between Model 1 and Model 2 is 3.35e-2 < 0.001, indicating a significant difference in fit.

2.4 ROC Curve

The ROC curve is a method for evaluating binary classification models with wide applications in signal detection theory. The ROC curve is plotted based on different classification thresholds, reflecting the relationship between True Positive Rate (TPR) and False Positive Rate (FPR) at different response thresholds (Bishop, 2006). In ROC curves, the x-axis represents the false positive rate, and the y-axis represents the true positive rate.

In ROC curves, TPR refers to the ratio of correctly classified positive cases to all actual positive cases. FPR refers to the ratio of negative cases incorrectly classified as positive to all actual negative cases. Here, positive cases are correct responses (i.e., signals in signal detection theory), while negative cases are incorrect responses (i.e., noise). To plot ROC curves, we need to vary response thresholds and calculate false positive rates and true positive rates at each threshold.

ROC curves demonstrate model performance at different response thresholds. AUC (Area Under Curve) measures the area under the ROC curve. AUC values range between 0 and 1, indicating the classifier's ability to distinguish between

positive and negative cases. An AUC of 0.5 represents random prediction, while values closer to 1 indicate better classifier performance. Generally, when AUC exceeds 0.8, we can consider the model's performance to be relatively good.

ROC curves perform well when positive and negative samples are balanced, but when sample sizes differ significantly, the Precision-Recall Curve (PRC) is a more appropriate metric (Davis & Goadrich, 2006).

2.5 Posterior Predictive Check

Posterior predictive check is typically not considered a goodness-of-fit metric, but because this method can also measure how well a model fits original data, this article treats it as a type of goodness-of-fit metric.

Posterior predictive check belongs to model validation methods, examining a model's ability to reproduce sample data (Palminteri et al., 2017; Steingroever et al., 2014; Vandekerckhove et al., 2011). The formula is:

$$p(y_{rep}|y, M) = \int p(y_{rep}|\theta, M)p(\theta|y, M)d\theta$$

where M is the model, y is the sample data, and y_{rep} is the sample data reproduced by the model (Gelman, Carlin, et al., 2013; Zhang et al., 2020).

In practical application, the posterior predictive check process is as follows: after fitting the model and obtaining fitted parameters, substitute these parameters into the model to generate simulated data. Then, through plotting or calculating statistical indices (such as MSE), compare differences between model-simulated data and real data to assess model fit and predictive ability (van de Schoot et al., 2021).

Posterior predictive checks can avoid problems that may arise from using only model comparison indices. For example, Palminteri et al. (2017) demonstrated through a simulation study that assuming two models A and B, even when model selection indices favor model A in most cases, model A might fail to simulate the overall trend of data variation while model B can. Therefore, beyond traditional goodness-of-fit metrics, simulating data is crucial for model evaluation.

Although posterior predictive check is a concept in Bayesian statistics, this does not mean it is only applicable to Bayesian parameter estimation. For non-Bayesian parameter estimation models, we can only obtain point estimates of parameters, but we can still use these point estimates to simulate data and compare them with real data. While posterior predictive checks have not been widely used in past computational modeling research, an increasing number of recent studies have adopted this method for model evaluation. It is foreseeable that in future research, posterior predictive checks may become an essential step (Zhang et al., 2020).

3 Cross-Validation-Based Metrics

Cross-validation is a fundamental method in machine learning for testing model generalization ability to out-of-sample data. However, in psychology, this method has only recently gained attention (Daniel et al., 2020; Verstynen & Kording, 2023). The cross-validation process involves first dividing the dataset into training and validation sets, then fitting different models on the training set, and finally comparing prediction accuracy of different models on the validation set to select the optimal model (Friedman et al., 2001; Geisser & Eddy, 1979). Notably, the goodness-of-fit indices introduced earlier are all used to validate model performance on the validation set.

Cross-validation has three main advantages. First, compared to many indices built on assumptions and derivations, cross-validation uses computational power to replace complex derivations, making it extremely simple and intuitive. Second, cross-validation naturally incorporates the three factors of model complexity (number of parameters, parameter space range, and mathematical form) when balancing model fit and complexity, a feature many indices lack. Third, cross-validation can serve not only as a relative metric for model selection but can also be combined with statistical indices such as MSE and AUC mentioned earlier to assess model fit to data distributions.

Common cross-validation methods include K-fold cross-validation and Leave-One-Out cross-validation (LOO-CV). K-fold cross-validation divides data into K folds, using K-1 folds as training data and the remaining fold as validation data. Leave-one-out cross-validation is a special case of K-fold cross-validation, where each sample is taken out as a test set while the remaining samples serve as the training set. For example, in a dataset with N samples, N-1 data samples serve as the training set while the remaining one sample is the validation set, meaning K = N. Leave-one-out cross-validation requires N evaluations to complete predictions for all data samples, making it computationally expensive. When sample data noise is minimal, leave-one-out can achieve performance at least as good as any K-value K-fold cross-validation; when sample data noise is substantial, leave-one-out has larger generalization error (Zhang & Yang, 2015).

Although cross-validation is the most commonly used method for validating model generalization ability in machine learning, its use is not widespread in cognitive modeling, primarily because leave-one-out cross-validation is often computationally expensive, while K-fold cross-validation faces the question of how many folds to use. Considering limitations in sample size and computational complexity, cognitive modeling researchers often use information criterion approximations instead of cross-validation indices. This article introduces four common such indices: AIC, DIC, WAIC, and PSIS-Loo-CV.

3.1 AIC

AIC (Akaike Information Criterion) is one of the earliest model comparison indices (Akaike, 1974) with solid theoretical foundations. First, AIC approxi-

mates the KL divergence between the data distribution predicted by the model and the true data distribution. Second, AIC has been shown to asymptotically approximate out-of-sample predictive accuracy and LOO-CV (Stone, 1977).

The AIC calculation formula is:

$$AIC = -2 \times \log L(\hat{\theta}|y) + 2 \times K$$

where $\log L(\hat{\theta}|y)$ is the log likelihood value at the optimal parameters $\hat{\theta}$ obtained through maximum likelihood estimation or maximum a posteriori estimation (see Section 0), and K is the number of parameters, serving as a penalty for model complexity. Smaller AIC values indicate better model fit.

AIC may perform poorly with small sample sizes (Sugiura, 1978), leading researchers to propose AICc (Hurvich & Tsai, 1989) with small-sample bias correction. The AICc calculation formula is:

$$AICc = -2 \times \log L(\hat{\theta}|y) + 2 \times K \times \frac{n}{n-K-1} = AIC + \frac{2K(K+1)}{n-K-1}$$

where n represents the number of trials. AICc converges to AIC with large sample sizes. With small sample sizes, AICc penalizes complex models more heavily than AIC. In cognitive modeling, due to the limited number of trials subjects complete in behavioral experiments, AICc is often a more appropriate index than AIC (Li et al., 2020; Li & Ma, 2021; Suzuki et al., 2012).

Regarding how large an AIC difference must be to demonstrate that one model is superior to another, Burnham and Anderson (2004) suggest that when the absolute difference in AIC between two models is less than 2, the two models are virtually indistinguishable; when the difference is between 4 and 7, there is modest evidence supporting the model with the smaller AIC; when the difference exceeds 10, there is strong evidence that the model with the smaller AIC is optimal. Additionally, since AIC asymptotically follows a chi-square distribution (Anderson & Burnham, 2004), researchers can use chi-square tests to compare whether AIC values differ significantly between models.

AIC can also be used to calculate Akaike weights (Wagenmakers & Farrell, 2004). Assuming there are M models, the Akaike weight for the jth model is calculated as:

$$\Delta AIC_j = AIC_j - \min AIC$$

$$w_j = \frac{\exp(-0.5 \times \Delta AIC_j)}{\sum_{i=1}^{M} \exp(-0.5 \times \Delta AIC_i)}$$

The first formula represents the difference between each model and the best model, and these differences are mapped to the 0-1 interval through the second formula, representing the weights of different models. The second formula is called the softmax formula, where ΔAIC multiplied by -0.5 ensures that models with smaller AIC receive higher weights. Anderson and Burnham (2004) consider Akaike weights an approximation of posterior model probability (PMP) $p(M|y)$, representing the probability that a model is the best among candidate models given the sample data.

AIC is widely used in cognitive modeling but has several drawbacks. First, as an approximation of out-of-sample predictive ability, AIC is less accurate than indices such as WAIC and PSIS-Loo-CV that will be introduced later. Second, AIC uses plug-in predictive probability $p(y_{rep}|\hat{\theta})$ in its derivation to assess in-sample predictive accuracy rather than evaluating the complete predictive distribution, leading to some bias in out-of-sample predictions. Finally, when measuring model complexity, AIC only considers the number of parameters, ignoring the other two factors affecting model complexity summarized by Myung and Pitt (1997).

3.2 DIC

DIC (Deviance Information Criterion) is one of the most common model selection indices in Bayesian statistics, with its theory based on expected log pointwise predictive density for a new dataset (elpd). DIC approximates elpd and is therefore only applicable to Bayesian parameter estimation models. Bayesian parameter approximation typically has two implementation approaches: sampling approximation methods based primarily on Markov Chain Monte Carlo (MCMC), and approximation methods such as Variational Inference (VI) that solve through approximating the posterior distribution. Sampling approximation methods are more computationally intensive and slower but typically yield more accurate results. DIC calculation requires posterior samples obtained from MCMC.

DIC is often considered the Bayesian parameter estimation version of AIC, but unlike AIC, DIC is only applicable to models estimated via MCMC sampling (Spiegelhalter et al., 2002).

The DIC calculation formula is $DIC = -2D(\bar{\theta}) + 2 \times p_D$. Here, $\bar{\theta}$ is the mean of the parameter posterior distribution, and $D(\theta)$ is the deviance between real data and model-predicted distribution, measuring model performance. The deviance formula is:

$$D(\theta_s) = -2 \times \log L(y|\theta_s)$$

where s represents MCMC samples, so θ_s are parameter values from MCMC samples. The first term of the DIC formula is -2 times the deviance at the mean of the parameter posterior distribution, representing model fit. The second

term, p_D, is called the effective number of parameters, serving as a penalty term for model complexity, calculated as:

$$p_D = \bar{D}(\theta) - D(\bar{\theta})$$

$$\bar{D}(\theta) = -2 \times \left(\frac{\sum \log L(y|\theta_s)}{S} \right)$$

Additionally, Gelman, Carlin, et al. (2013) proposed using the variance of deviance as the effective number of parameters:

$$p_D = 0.5 \times \mathrm{Var}(\log L(y|\theta))$$

Like AIC, smaller DIC values indicate better model fit. When we divide DIC by -2, we obtain DIC's approximation of elpd. Unlike AIC, p_D in DIC considers not only the number of model parameters but is also sensitive to other factors affecting model complexity summarized by Myung and Pitt (1997). Because of this characteristic, DIC often provides researchers with more insights. For example, the LBA (Linear Ballistic Accumulator) model and DDM (Drift-Diffusion Model) both belong to the class of sequential sampling models for reaction times (Brown & Heathcote, 2008). LBA is generally considered a simplified version of DDM. To verify which of these is more complex, Donkin et al. compared them using DIC (Donkin et al., 2009). The results showed that despite LBA having fewer parameters than the drift-diffusion model, LBA had a larger p_D in DIC, suggesting that LBA may not actually simplify DDM.

Compared to AIC, DIC provides a more accurate approximation of out-of-sample predictive ability. However, DIC has several problems. First, DIC's performance is heavily influenced by the shape of the parameter posterior distribution and the stability of parameter point estimates. Second, when point estimates of the parameter posterior distribution cannot be well-represented by the mean, or when model parameters are not from exponential family distributions, DIC estimates may be biased. For example, when the parameter posterior distribution is multimodal, DIC is prone to being less than 0 (Evans et al., 2020; Spiegelhalter et al., 2014).

3.3 WAIC and PSIS-Loo-CV

WAIC (Widely Applicable Information Criterion) (Watanabe, 2010) and PSIS-Loo-CV (Pareto Smoothed Importance Sampling-Leave-One-Out Cross-Validation) (Vehtari et al., 2017) are similar to DIC in that they approximate elpd and are also only applicable to Bayesian models based on MCMC sampling.

Unlike DIC, WAIC uses lpd (Log Pointwise Predictive Density, also abbreviated as *lpd* in some articles):

$$\hat{lpd} = \sum \log\left(\frac{1}{S}\sum p(y_i|\theta_s)\right)$$

where i is the ith sample data point and S is the number of samples from the MCMC posterior distribution. Approximating elpd with lpd often overestimates elpd—that is, it overestimates model predictive ability. Therefore, WAIC introduces a correction term $\hat{p}_{waic}$ when calculating elpd. This term is similar to the number of parameters in AIC and p_D in DIC, serving to penalize model complexity. $\hat{p}_{waic}$ represents the estimated effective number of parameters, calculated as:

$$\hat{p}_{waic} = \sum \mathrm{Var}_{s=1}^{S}(\log p(y_i|\theta_s))$$

$$\widehat{elpd}_{waic} = \hat{lpd} - \hat{p}_{waic}$$

To make WAIC asymptotically follow a chi-square distribution, we can multiply it by -2. Notably, larger $\widehat{elpd}_{waic}$ indicates better out-of-sample predictive ability, while smaller WAIC indicates better model fit.

Compared to DIC, although WAIC also uses plug-in prediction to evaluate out-of-sample generalization ability, WAIC has several additional advantages. First, WAIC uses the entire posterior distribution to calculate the complexity penalty term, yielding more stable results. Second, WAIC performs better than DIC for models with non-Gaussian parameter posterior distributions (Myung & Pitt, 2018).

Bayesian leave-one-out cross-validation can also be used to approximate elpd. Its calculation formula is:

$$\widehat{elpd}_{loo} = \sum \log p(y_i|y_{-i})$$

$$p(y_i|y_{-i}) = \int p(y_i|\theta) \times p(\theta|y_{-i})d\theta$$

where i represents the ith data sample point. The information criterion based on $\widehat{elpd}_{loo}$ is LOOIC (Leave-One-Out Cross-Validation Information Criterion), which is $\widehat{elpd}_{loo}$ multiplied by -2. For leave-one-out cross-validation, the penalty term for model complexity is the difference between $\widehat{elpd}_{loo}$ and $\hat{lpd}$.

Bayesian leave-one-out cross-validation is extremely computationally expensive. To simplify calculation, Vehtari et al. (2017) proposed PSIS-Loo-CV to approximate full LOO-CV. PSIS-Loo-CV uses MCMC samples, significantly reducing computational load. Because the R package `loo` incorporates this algorithm, it has been widely applied in practical research. Additionally, PSIS-Loo-CV

provides a model diagnostic metric: the k value of the Pareto distribution. If the k values for the vast majority of data points exceed 0.7, it suggests potential problems with model specification.

Beyond using WAIC and PSIS-Loo-CV for model comparison, Vehtari et al. (2019) also recommend combining PSIS-Loo-CV with stacking methods from ensemble learning (Friedman et al., 2001) to calculate weights for each model, with details available in Yao et al. (2018). Like Akaike weights, model weights from stacking methods can be used for model averaging. Notably, when stacking weights are used for model comparison, models with similar performance will "share" weights, resulting in lower and similar weights for both (Sivula et al., 2020).

Compared to WAIC, PSIS-Loo-CV has been shown to be a better approximation of elpd (Vehtari et al., 2016), enabling PSIS-Loo-CV to more comprehensively consider the three factors affecting model complexity proposed by Myung and Pitt (1997). Moreover, the R package `loo` developed by Vehtari et al. (2017) lowers the barrier to use—researchers only need to input the likelihood function from MCMC sampling to calculate WAIC and PSIS-Loo-CV. Specific recommendations for using WAIC and PSIS-Loo-CV can be found in Vehtari (2022).

3.4 Summary of Different Cross-Validation Approximation Indices

Table 2. Advantages, disadvantages, and applicable parameter estimation ranges of various cross-validation approximation indices

Metric	Applicable Parameter Estimation Methods	Advantages	Disadvantages
AIC	Maximum likelihood, MAP, Bayesian estimation	Simple to compute, usable with any parameter estimation method	Approximation accuracy for cross-validation is not as good as the latter three
DIC	Bayesian parameter estimation	Simple to compute, provided by most Bayesian statistical software	Does not utilize the entire parameter posterior distribution obtained from Bayesian estimation

Metric	Applicable Parameter Estimation Methods	Advantages	Disadvantages
WAIC	Bayesian parameter estimation	More accurate approximation of cross-validation	Can be affected by MCMC sampling extreme values
PSIS-Loo-CV	Bayesian parameter estimation	More accurate approximation of cross-validation	Can be affected by MCMC sampling extreme values

Cross-validation-based metrics are widely used in cognitive modeling. With the recent popularity of black-box MCMC software enabling researchers to easily use Bayesian parameter estimation, this has greatly promoted the use of DIC, WAIC, and LOO-CV.

Although these indices are built on different assumptions and approximation methods—AIC is more applied to models fitted with maximum likelihood or MAP estimation, while DIC, WAIC, and PSIS-Loo-CV are used for Bayesian parameter estimation models with MCMC—in some cognitive modeling applications, their differences are not pronounced. For example, Evans (2019) compared AIC, DIC, and WAIC on LBA models and found similar performance, though DIC and WAIC performed slightly better than AIC. Similarly, Westbrook et al. (2020) used AIC and DIC to compare different attentional drift-diffusion models (aDDM), and the results were nearly identical.

4 Marginal Likelihood

Marginal likelihood, also called model evidence, is another major class of model evaluation metrics and the core of Bayesian Model Selection (BMS). The Bayesian parameter estimation formula is:

$$p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{\int p(y|\theta) \times p(\theta) d\theta}$$

The left side $p(\theta|y)$ is the parameter posterior distribution, the first term on the right $p(\theta)$ is the parameter prior distribution, and the second term $p(y|\theta)$ is the likelihood function. The problem with this formula is that it ignores the model M term. If we modify this formula to include M:

$$p(\theta|y, M) = \frac{p(y|\theta, M) \times p(\theta, M)}{\int p(y|\theta, M) \times p(\theta, M) d\theta}$$

The denominator in this Bayesian formula is then the model's marginal likelihood or model evidence. The larger the marginal likelihood, the better the model explains the sample data.

Marginal likelihood can balance model complexity and fit. For example, simpler models may have lower goodness-of-fit but higher marginal likelihood because they have less parameter space uncertainty. Conversely, complex models may have higher goodness-of-fit but lower marginal likelihood because they have greater parameter space uncertainty (MacKay, 2003).

Marginal likelihood simultaneously considers the three factors affecting model complexity summarized by Myung and Pitt (1997), as shown in the figure. Models that are too simple often assign low probability $p(M|y)$ to observed data, resulting in small marginal likelihood; models that are too complex have broader data distributions but also assign low probability $p(M|y)$ to current observed data, resulting in small marginal likelihood; only when complexity is moderate will the marginal likelihood corresponding to observed data be large.

Figure 3. Marginal likelihood penalty for different types of models. The x-axis represents data values; the y-axis represents likelihood values corresponding to data values.

Marginal likelihood is also particularly sensitive to prior information in Bayesian parameter fitting. For example, when using weakly informative prior distributions, complex models have smaller marginal likelihood than simple models; when using narrower, more informative prior distributions, complex models may have larger marginal likelihood than simple models (Farrell & Lewandowsky, 2018).

Marginal likelihood has two main practical problems. First, prior distributions significantly impact marginal likelihood calculation results. Inappropriate prior distributions, especially with many data points, can substantially affect parameter estimation results and consequently marginal likelihood calculations (Boehm et al., 2018). Regarding prior selection, subjective Bayesian approaches argue that prior distributions should be chosen based on existing knowledge and beliefs, while objective Bayesian approaches attempt to eliminate personal factors in prior selection and use non-informative prior distributions such as Jeffreys default prior distribution (Jeffreys, 1998; Vandekerckhove et al., 2015). To select more appropriate prior distributions, researchers can use prior sensitivity analysis to examine the impact of different prior distributions on marginal likelihood.

The second problem is that calculating marginal likelihood requires integrating the product of the prior distribution and model likelihood function over the

entire parameter space. However, only a few simple models have marginal likelihoods that can be solved directly; for most models, marginal likelihood cannot be easily calculated. Therefore, many approximation and sampling integration methods have been proposed for calculating marginal likelihood. Several common methods are introduced below.

4.1 BIC

BIC (Bayesian Information Criterion) (Schwarz, 1978), similar to AIC, is one of the most classical and widely used model selection indices. BIC is a special case of marginal likelihood with Laplace approximation (Bishop, 2006). When calculating Laplace approximation, assuming a non-informative prior and when the number of data points is extremely large, according to the law of large numbers, the Laplace approximation result can be simplified to BIC. Although BIC is based on Bayesian model comparison, because of its computational simplicity, it is also commonly used for maximum likelihood estimation that does not consider prior effects.

The BIC calculation formula is:

$$BIC = -2 \times \log L(\hat{\theta}|y) + K \times \ln(n)$$

where $K \ln(n)$ is the complexity penalty term in BIC, K is the number of parameters, and n is the number of trials. Thus, BIC considers not only the impact of parameter count on complexity penalty but also uses data size as a key factor in penalizing model complexity. Like AIC, smaller BIC values indicate better model fit. Additionally, there is sample-adjusted BIC (SABIC) (Sclove, 1987), though it lacks theoretical justification and is rarely used (Dziak et al., 2020).

Although BIC is the most common model selection index (Wilson & Collins, 2019), it still has drawbacks. First, BIC's penalty for model complexity only considers parameter count and sample size, ignoring the other two factors affecting model complexity summarized by Myung and Pitt (1997): parameter space range and mathematical form. Second, although BIC is derived within the Bayesian theoretical framework, it does not consider the impact of different prior information on results.

4.2 Approximation Methods for Marginal Likelihood

Approximation methods for marginal likelihood introduced in this article include Savage-Dickey Ratio (SDR), Laplace approximation, Kernel Density Estimation (KDE), and Variational Inference. Compared to BIC, these methods consider prior distribution effects without significantly increasing computational load; compared to sampling methods introduced later, approximation methods have larger errors but are far less computationally demanding, making them applicable in many studies.

Savage-Dickey Ratio is suitable for calculating Bayes factors between nested models in model comparison (Dickey, 1973; Dickey, 1976; Wagenmakers et al., 2010). Assuming the parameter missing from the simple model is θ, Savage-Dickey Ratio simplifies the Bayes factor calculation for nested models to the ratio of posterior probability to prior probability when θ equals 0 in the full model:

$$BF_{01} = \frac{\pi(\theta_0|y, M)}{\pi(\theta_0|M)}$$

The Savage-Dickey problem is that it is only suitable when parameters have low collinearity, whereas parameters in many cognitive models often have some degree of collinearity (Heck, 2019).

Laplace approximation is mainly applied when fitting models using maximum a posteriori estimation. Its main idea is to use a multivariate Gaussian distribution to approximate the parameter distribution and use Taylor expansion to avoid integration problems. Compared to BIC, Laplace approximation marginal likelihood considers prior distribution effects and has smaller computational error. The Laplace approximation formula for marginal likelihood is:

$$\log p(y|M) \approx \log L(\hat{\theta}|y) + \log p(\hat{\theta}|M) + \frac{d}{2} \times \log 2\pi - \frac{1}{2} \log |H|$$

where $|H|$ is the determinant of the Hessian matrix of the negative log posterior. Laplace approximation is one of the most common methods for approximately calculating marginal likelihood in psychology (Gershman, 2016; Huys et al., 2011; Myung & Pitt, 1997). The key step is calculating the determinant of the Hessian matrix, but when the Hessian matrix is not positive definite, the $\log |H|$ term may be NaN.

Kernel density estimation methods can calculate marginal likelihood using parameter posterior distributions obtained from MCMC sampling. Kernel density estimation uses non-parametric statistical methods to calculate the posterior probability of parameters $p(\hat{\theta}|y) = k(\hat{\theta}|\theta, \phi)$. Here, k is the density kernel function, typically Gaussian (Wasserman, 2006), and ϕ is the bandwidth of the density kernel. θ represents parameter samples obtained from MCMC sampling, while $\hat{\theta}$ is the point estimate representative of the MCMC sampling distribution, generally the point with highest probability density.

After obtaining the parameter posterior probability $p(\hat{\theta}|y)$, according to Bayes' formula, we can directly obtain marginal likelihood:

$$p(y|M) = \frac{L(\hat{\theta}|y) \times p(\theta|M)}{p(\hat{\theta}|y)}$$

Kernel density estimation methods are simple to compute and not constrained by the Hessian matrix. Some simulation studies have found that its performance is better than Laplace approximation and other methods (Bos, 2002).

Variational Inference is another common Bayesian parameter estimation method besides sampling methods. Unlike sampling methods, Variational Inference attempts to approximate the parameter posterior distribution $p(\theta|D)$ with a variational distribution $q(z)$, thereby transforming the integration problem in Bayes' formula into an optimization problem (Bishop, 2006). Variational Inference has many applications not only in Bayesian parameter estimation but can also serve as a theory for understanding cognitive processes (Friston et al., 2006). The optimization function in Variational Inference is called ELBO (Evidence Lower Bound) or Negative Free Energy (Bishop, 2006; Friston et al., 2007), which is a lower bound of the log marginal likelihood. Maximizing ELBO yields an estimate of marginal likelihood. The ELBO formula is:

$$ELBO = E_{q(z)}[\log p(\theta, y|M)] = E_{q(z)}[\log p(y|\theta, M)] + D_{KL}(q(z)||p(\theta|M))$$

The ELBO formula shows that marginal likelihood can be divided into two parts: the first part is the expected value of the likelihood function under the variational distribution, representing model fit; the second part is the KL divergence between the variational distribution and prior distribution, representing the difference between posterior and prior. When model fit is worse or the difference between prior and posterior distributions is larger, marginal likelihood is smaller (Stephan et al., 2009).

In practical applications, the Variational Inference toolbox VBA based on Matlab can return optimized ELBO after model fitting (Daunizeau et al., 2014). Additionally, models fitted with Stan also return unstandardized posterior distribution probabilities and variational distribution probabilities that can be used to calculate ELBO. The problem with Variational Inference methods is that they provide a lower bound of marginal likelihood, and few theoretical studies have focused on the approximation error of ELBO to marginal likelihood (Blei et al., 2017).

4.3 Sampling Methods for Marginal Likelihood

Monte Carlo sampling methods are common statistical simulation methods. When an integral formula is difficult to solve directly, we can continuously sample numerically, substitute into the formula for calculation, and gradually approach the integral result. Because the marginal likelihood integral for complex models cannot be solved analytically, many Monte Carlo sampling algorithms have been applied to calculate marginal likelihood.

Sampling methods are numerous, including Thermodynamic Integration, Sequential Monte Carlo sampler (SMC), and particle MCMC methods. However,

due to the lack of user-friendly software, these methods have limited application (Doucet & Johansen, 2009; Murphy, 2023). In contrast, Importance Sampling (Gamerman & Lopes, 2006; Hammersley, 2013) and Bridge Sampling (Gronau et al., 2017; Meng & Wong, 1996) have user-friendly software or are computationally simple, making them widely used in psychological research. Notably, these two sampling methods are not identical to MCMC for model fitting. They are more applied to numerical integration, while MCMC is mainly used for parameter fitting.

Importance Sampling belongs to Monte Carlo methods, with its key being the introduction of an importance sampling distribution. When sampling from a distribution is difficult or yields low-quality samples, we can settle for sampling from the importance distribution (Bishop, 2006). When calculating marginal likelihood, we first introduce the importance sampling distribution $g_{IS}(\theta)$, obtaining:

$$p(y|M) = \int p(y|\theta, M) \times p(\theta|M) d\theta = \int \frac{p(y|\theta, M) \times p(\theta|M)}{g_{IS}(\theta)} \times g_{IS}(\theta) d\theta = E_{g_{IS}(\theta)} \left(\frac{p(y|\theta, M) \times p(\theta|M)}{g_{IS}(\theta)} \right)$$

Therefore, marginal likelihood can be obtained by:

$$\hat{p}(y|M) = \sum \frac{p(y|\theta_i, M) \times p(\theta_i|M)}{g_{IS}(\theta)}, \quad \tilde{\theta}_i \sim g_{IS}(\theta)$$

By continuously sampling from the importance distribution, substituting into Bayes' formula for calculation, and summing results from different samples, we can obtain marginal likelihood. In importance sampling, the choice of importance distribution greatly affects results. To ensure estimates have small variance, $g_{IS}(\theta)$ is typically a distribution with thick tails. Additionally, when using importance sampling to calculate the reciprocal of marginal likelihood $1/\hat{p}(y|M)$, this importance sampling is also called RIS (Reverse Importance Sampling) (Gelfand & Dey, 1994). Conversely, RIS sampling distributions require distributions with thinner tails.

Using MCMC sampling to obtain parameter posterior samples for calculating marginal likelihood can significantly reduce computational load. This importance sampling is called the Harmonic Mean Estimator. The harmonic mean estimator is easy to compute but has large variance in its results.

Common methods to improve harmonic mean estimator performance include: First, using Weighted Importance Sampling (Acerbi et al., 2018). This method requires multiplying RIS by a function $f(\theta)$ with thin tails, where $\int f(\theta) d\theta = 1$, so $f(\theta)$ can be a multivariate Gaussian distribution. The RIS calculation formula is:

$$\widehat{p}(y|M) = \frac{f(\theta_i)}{p(y|\theta_i, M) \times p(\theta_i|M)}$$

Second, replacing MCMC samples with a mixture distribution of uniform or Gaussian distributions and MCMC samples (Steingroever et al., 2016; Vandekerckhove et al., 2015). This method is convenient to compute and has many applications in psychology.

Bridge Sampling is an improvement over Importance Sampling. Like Importance Sampling, Bridge Sampling also uses MCMC samples. Compared to the simpler Importance Sampling, Bridge Sampling avoids the step of selecting a distribution, yields smaller variance in results, and is more suitable for hierarchical models. Bridge Sampling's characteristic is that by introducing a bridge distribution connecting the target distribution and proposal distribution, it reduces the variance of marginal likelihood calculation and improves computational accuracy (Meng & Wong, 1996). Bridge Sampling's disadvantage is that its calculation is relatively complex, requiring iterative computation until results stabilize, which increases time and resource requirements, as detailed in Gronau et al. (2017). The R package `bridgesampling` developed by Gronau et al. simplifies the calculation process, allowing models fitted with JAGS and Stan to use this package for marginal likelihood calculation.

4.4 Summary of Different Marginal Likelihood Calculation Methods

Table 3. Advantages, disadvantages, and applicable parameter estimation ranges of various marginal likelihood approximation indices

Method	Applicable Parameter Estimation Methods	Advantages	Disadvantages
BIC	Maximum likelihood, MAP, Bayesian estimation	Simple to compute, usable with any parameter estimation method	No prior influence, poorer approximation of marginal likelihood than the latter four methods
Savage-Dickey Ratio	Bayesian parameter estimation	Simpler to compute than sampling methods	Few studies use it; no tool package, requires manual implementation by researchers

Method	Applicable Parameter Estimation Methods	Advantages	Disadvantages
Laplace Approximation	Maximum likelihood, MAP, Bayesian estimation	Usable with any parameter estimation method	Hessian matrix may be NaN; no tool package, requires manual implementation
Importance Sampling	Bayesian parameter estimation	Simpler to compute than Bridge Sampling	Easily affected by MCMC sampling extreme values
Bridge Sampling	Bayesian parameter estimation	More accurate approximation of marginal likelihood than Importance Sampling	Complex computation steps; only R package `bridgesampling` provides convenient implementation

There are many methods for calculating marginal likelihood, and the choice depends on the specific application context. BIC is the simplest method but also has the largest error. Additionally, because BIC approximates marginal likelihood with non-informative priors, theoretically using BIC will tend to select simpler models. Evans (2019) suggests that using BIC is inappropriate when researchers fit models with informative prior distributions. The prior distribution for calculating marginal likelihood should be consistent with the prior used for model fitting.

Table 3 summarizes the advantages and disadvantages of each marginal likelihood index. When using MAP estimation for model fitting, Laplace approximation is a simpler method. If using MCMC sampling and the model is not hierarchical, Importance Sampling, Laplace approximation, or KDE methods are more appropriate because they are less computationally demanding. If the model is hierarchical, where the Hessian matrix determinant is not easily computed for Laplace approximation and Importance Sampling faces difficulties in selecting sampling distributions, Bridge Sampling becomes a more reasonable choice.

5 Case Study of Model Comparison Calculations

The previous sections introduced commonly used model comparison indices in cognitive modeling. The following sections use the orthogonal Go/No-Go paradigm as an example to introduce the specific calculation and usage methods of some common model indices (Cavanagh et al., 2013; Dorfman & Gershman, 2019; Guitart-Masip et al., 2012). The example data were simulated using cognitive models described below. Simulated data and subsequent model comparison index calculations were performed using R, with specific code available in the online materials: https://github.com/zaizibai/model_{comparison}.

The orthogonal Go/No-Go paradigm is commonly used to study the relationship between Pavlovian and instrumental learning. Figure 4 [Figure 4: see original paper] shows the basic procedure of this paradigm. It is a 2$×$2 within-subject experimental design where the first variable is stimulus-response action: Go and No-Go; the second variable is feedback type after behavioral response: reward gain and punishment avoidance. The combination of stimulus-response action and feedback type creates four experimental conditions: Go-to-gain-reward, Go-to-avoid-punishment, No-Go-to-gain-reward, and No-Go-to-avoid-punishment. Feedback in each condition is not 100% deterministic. In the "Go-to-avoid-punishment" condition, a correct response (i.e., Go) has an 80% probability of avoiding punishment but a 20% probability of failing to avoid; an incorrect response (i.e., No-Go) has an 80% probability of receiving punishment and a 20% probability of avoiding it. The image shown on the first screen at trial onset is called the cue, with four types corresponding one-to-one with experimental conditions. At the experiment's start, subjects do not know the correct response for each condition and must learn it through feedback. According to learning theory, in this paradigm, people tend to have Go responses when feedback is reward gain, and No-Go responses when feedback is punishment avoidance (Dayan et al., 2006).

Figure 4. Experimental design of the case study, adapted from Betts et al. (2020). The procedure for a single trial is as follows: subjects first see a cue, and after the cue disappears, they must make a Go or No-Go response. After responding, the screen displays the outcome. In this task, subjects need to actively learn the correct response for different cues and whether correct outcomes avoid punishment or gain reward.

Researchers typically use simple reinforcement learning models to model data from this paradigm. These models posit that human decision-making is influenced by two learning factors: Pavlovian learning and instrumental learning. Instrumental learning originates from Skinner's theory of instrumental learning, forming stimulus-response-outcome (SRO) associations, while Pavlovian learning forms stimulus-outcome associations independent of response. Specifically, the decision weight for choosing Go or No-Go responses is calculated as follows:

$$w = b + Q + \pi \times V$$

where b represents an individual's natural preference for Go or No-Go responses, called the Go bias parameter; Q is the instrumental learning decision variable; V is the Pavlovian effect decision variable; and π is its scaling parameter. For specific details about this model, see Betts et al. (2020) or Swart et al. (2017).

This article uses publicly available data from Raab & Hartley (2020), available at: https://osf.io/4h6ne/. This dataset includes 61 subjects. Figure 5 [Figure 5: see original paper] presents the raw data for Go response choices across the four conditions. For this data, we fitted four models in total: the full model using Equation 1 (Model 1), a model without Pavlovian effects and Go bias parameters (Model 2), a model without Pavlovian effects but with Go bias parameters (Model 3), and a model without Go bias parameters but with Pavlovian effects (Model 4). Notably, we fitted these four models using both point-estimate MAP estimation and hierarchical Bayesian parameter estimation. Hierarchical Bayesian parameter estimation was implemented using the probabilistic programming software Stan (Carpenter et al., 2017).

Figure 5. Trial-by-trial behavioral data for the case study. The x-axis shows trial number, and the y-axis shows the proportion of Go responses. Four colors represent the four cues.

As trial number increases, individual behavior gradually stabilizes, reflecting the effect of instrumental learning. The asymmetry in the proportion of Go responses under reward-gain and punishment-avoidance cues reflects Pavlovian effects. Specifically, individuals are more likely to make Go responses to gain rewards but more likely to make No-Go responses to avoid punishment.

5.1 Calculation of Goodness-of-Fit Metrics

The case study data in this article are discrete choice variables, so we can calculate likelihood functions, pseudo-r^2, and ROC curve indices. In the case study, we only use pseudo-r^2 as an example. According to Equation 7, we calculated pseudo-r^2 for the four models. Results show Model 1 pseudo-$r^2 = 0.157$, Model 2 pseudo-$r^2 = 0.132$, Model 3 pseudo-$r^2 = 0.147$, and Model 4 pseudo-$r^2 = 0.139$. This indicates that Model 1 has better absolute fit than the other models.

Although Model 1 has better absolute fit, goodness-of-fit does not consider model complexity. In the following two sections, we introduce the calculation and usage methods for cross-validation metrics and marginal likelihood metrics.

5.2 Calculation and Usage of Cross-Validation Metrics

The typical usage method for cross-validation metrics is to compare the mean or sum of indices across all subjects. However, Devine et al. (2023) found through simulation studies that methods considering uncertainty in model comparison indices significantly improve correct model comparison rates, while simply comparing mean index values tends to have high false positive rates. Devine et al. (2023) recommend using the method adopted by Vehtari et al. (2017), per-

forming Wald tests to compare different models based on Bayesian model metrics such as DIC, WAIC, and PSIS-Loo-CV. The Wald test procedure involves calculating the mean and standard error of differences in model metrics; if the mean exceeds 1.96 standard errors, the difference between models is judged significant. According to Vehtari et al. (2017), the standard error formula for a single model comparison index is:

$$se(elpd) = \sqrt{\frac{1}{N-1}\sum(elpd_i - \overline{elpd})^2}$$

where i is a sample data point, N in psychology experiments is all trials across all subjects, and $\overline{elpd}$ is the mean of $elpd_i$. Similarly, when calculating the standard error of the difference between two model comparison indices, first calculate the difference in model comparison indices at each data point, then calculate the standard error of the N differences:

$$se(elpd_A - elpd_B) = \sqrt{\frac{1}{N-1}\sum\left((elpd_{Ai} - elpd_{Bi}) - \overline{(elpd_A - elpd_B)}\right)^2}$$

where $\overline{(elpd_A - elpd_B)}$ is the mean difference between the two model comparison indices. The Wald test incorporates uncertainty in model indices and has lower false positive probability.

Figure 6. Evaluation of four models by different cross-validation approximation indices. Smaller information criterion values indicate better model fit.

Note: Results from PSIS-Loo-CV are often denoted as LOOIC (Leave-One-Out Information Criterion).

In the case study, we calculated AIC using MAP estimation results and calculated DIC, WAIC, and PSIS-Loo-CV using hierarchical Bayesian parameter estimation results, as shown in the figure. The results from different indices are consistent, with Models 1 and 3 performing better than the other two.

In the case study, we performed Wald tests comparing Model 1 and Model 3. Results show significant differences in DIC, WAIC, and LOO-CV between the two models, all favoring Model 1 over Model 3, with similar results across the three: $\Delta DIC = 48.23 > 1.96 \times \sigma_{DIC} = 22.52$, $\Delta WAIC = 44.3 > 1.96 \times \sigma_{WAIC} = 22.15$, $\Delta PSIS-LOO-CV = 38.5 > 1.96 \times \sigma_{PSIS-LOO-CV} = 21.77$. Here, Δ represents the difference in cross-validation metrics between Model 2 and Model 1, and σ is the standard error of model differences.

5.3 Calculation and Usage of Marginal Likelihood Metrics

As the core of Bayesian model comparison, marginal likelihood has many usage methods. The most common one is that when researchers compare two models,

they can calculate the ratio of marginal likelihoods for the two models, resulting in the Bayes factor (Kass & Raftery, 1995). The Bayes factor's characteristic is its ability to provide evidence for the null hypothesis, giving it many applications in current psychological research. For the use of Bayes factors in data analysis and interpretation of results, see 胡传鹏 et al. (2018). Additionally, BIC, as an approximation of marginal likelihood, can also be used to calculate Bayes factors and posterior model probabilities (Wagenmakers, 2007). The calculation method involves multiplying the difference in BIC between two models by -0.5, then transforming it through the exponential function into a Bayes factor:

$$BF_{10} = \exp\left(-\frac{BIC_1 - BIC_0}{2}\right)$$

Notably, unlike common data analysis, the two models compared by Bayes factors in cognitive modeling can be any two models as long as they model the same data. In contrast, the two models compared in t-tests and ANOVA must be alternative and null hypotheses.

In the case study, both BIC and Laplace approximation marginal likelihood are based on MAP estimation results, which we can use to calculate subject-level Bayes factors. In contrast, Bridge Sampling is suitable for hierarchical Bayesian estimation and can directly calculate group-level marginal likelihood values, thereby obtaining group-level Bayes factors (GBF).

Figure 7. Comparison of group-level marginal likelihood results calculated using BIC, Laplace approximation, and Bridge Sampling. Unlike cross-validation metrics, marginal likelihood does not always favor the most complex model (Model 1). When using BIC, the optimal model is the simplest Model 2; Laplace approximation supports Model 1; Bridge Sampling indicates Model 3 is optimal. This reflects that marginal likelihood has stronger penalization than cross-validation metrics. Moreover, differences in numerical values across indices are due not only to different approximation accuracies of marginal likelihood but also to differences in model fitting methods.

Figure 7. Evaluation of four models by different group marginal likelihood approximation indices. All indices are converted to log marginal likelihood, with larger values indicating better model fit.

Traditional model comparison typically selects a single optimal model, but a single model may both overfit and ignore model uncertainty. Researchers have proposed Bayesian model averaging, which simultaneously considers the weights of multiple models' influences to enhance the robustness of inferences based on models (Clyde et al., 2011; Hinne et al., 2020; Merlise & Edward, 2004). Through Bayesian model averaging, researchers can calculate inclusion Bayes factors ($BF_{inclusion}$) to compare different types of models:

$$BF_{inclusion} = \frac{p(M_{cat1}|y)}{p(M_{cat2}|y)} \times \frac{p(M_{cat2})}{p(M_{cat1})}$$

where $p(M_{cat1})$ and $p(M_{cat2})$ are the prior probabilities of Type 1 and Type 2 models, and $p(M_{cat1}|y)$ and $p(M_{cat2}|y)$ are the posterior probabilities of Type 1 and Type 2 models. The posterior probability calculation formula for models is:

$$p(M_i|y) = \frac{p(M_i|y) \times p(M_i)}{\sum p(M_k|y) \times p(M_k)}$$

$p(M|y)$ is marginal likelihood, and $p(M)$ is the prior probability of models, typically uniform distribution. $BF_{inclusion}$ compares different types of models combined, reducing the impact of model uncertainty.

Model averaging has wide applications in variable selection, meta-analysis, and other fields. For example, the ANOVA section in JASP uses Bayesian model averaging (王允宏 et al., 2022). However, its application in cognitive modeling is still limited. One of the few studies is Boehm et al. (2023), where the authors used model averaging to investigate the impact of speed-accuracy tradeoffs on DDM parameters and found that Bayesian model averaging can reduce the impact of model overfitting on DDM parameter estimation, making DDM parameter analysis results more accurate. Additionally, Bayesian model averaging is limited by marginal likelihood calculation; when marginal likelihood calculation is difficult, it is hard to calculate posterior model probabilities. A feasible method is to use Akaike weights or BIC to substitute for posterior model probabilities. Moreover, model weights combining stacking methods and PSIS-Loo-CV can also be used to substitute for posterior model probabilities (Yao et al., 2018).

In the case study, for convenience, we use BIC calculated earlier to compute $BF_{inclusion}$. We first calculated the BF and $BF_{inclusion}$ for models with and without the Go bias parameter. Models with parameter b include Models 1 and 4, while models without Go bias include Models 2 and 3. Traditional methods for determining whether Go bias significantly improves model performance typically compare Model 1 and Model 3, with the difference between these two models being only the inclusion of Go bias. The BF between Model 1 and Model 3 is 4.29, while the $BF_{inclusion}$ considering multiple models is 3.60, both supporting that Go bias does not significantly improve model fit. We used the same method to calculate BF and $BF_{inclusion}$ for models with and without Pavlovian effects, yielding $BF = 2.29$ and $BF_{inclusion} = 1.86$, also indicating that adding Pavlovian effects does not improve model fit. The results from model averaging methods are similar to traditional methods because in this case study, parameters in the simulated data are not correlated; however, when correlations between parameters are large, differences between BF and $BF_{inclusion}$ results will be substantial.

6 Summary and Outlook

Computational modeling has become increasingly widespread in experimental psychology research over the past decade, and model comparison is a critical component of cognitive modeling. Inappropriate model comparison may lead researchers to draw incorrect conclusions. Therefore, the rational use of model comparison indices is crucial for research based on computational models. This article reviews and summarizes common and emerging model selection indices in cognitive modeling, compares the two most common classes of indices—cross-validation-based and marginal likelihood-based—and recommends conditions for using different indices. Combined with a simple case study, we provide specific calculation examples.

Notably, many past studies using computational models have employed relatively simple model comparison indices such as AIC and BIC. While these indices have many advantages, they neglect many important factors affecting model complexity. In recent years, promoted indices such as WAIC, and marginal likelihood calculated by approximation/sampling methods, provide more comprehensive consideration of model complexity, making model comparison results based on these indices more stable and reliable. With the development of increasingly mature and easy-to-use tools, these indices will be more widely applied in research.

Furthermore, early cognitive modeling research mostly focused on using relative indices to evaluate model quality, neglecting the absolute goodness of model fit. This leads to a dilemma: even when we select an optimal model, it may not necessarily provide a complete description of sample data. Therefore, when conducting model comparison, we should first select the optimal model through relative indices, then evaluate the absolute goodness of model fit to current data through goodness-of-fit indices. Only when a model outperforms other candidate models on relative indices and has good absolute goodness-of-fit can it be considered the optimal model. With the popularization of methods such as posterior predictive checks, future research should increasingly combine relative and absolute indices for model evaluation and validation.

6.1 Debate Between Marginal Likelihood and Cross-Validation

This article has focused on introducing marginal likelihood and cross-validation, the two most common model comparison methods. Although they are based on very different theories, research has shown many connections between them. For example, Fong and Holmes (2020) proved that marginal likelihood is equivalent to cross-validation under certain specific conditions. However, which of these two is more suitable for practical research and how to choose between them remain controversial.

Modeling typically involves two scenarios: M-Closed and M-Open. The M-Closed scenario assumes that a “true” model exists among candidate models that can perfectly describe the data generation process. The M-Open scenario

assumes that no candidate model can perfectly describe the data generation process. In M-Open scenarios, the goal of model selection is to find the model that performs best among all candidate models, not to find the true model (Burnham & Anderson, 2004; Gelman, Hwang, et al., 2013).

If in an M-Closed scenario with nearly infinite data, marginal likelihood can select the "true" model. In M-Open scenarios, cross-validation is more suitable as it can find the model with the smallest KL divergence from the "true" model. Although cross-validation can also find the model with smallest KL divergence from data in M-Closed environments, it cannot identify the "true" model. Research shows that the advantages of marginal likelihood and cross-validation cannot be combined (Vrieze, 2012; Yang, 2005).

Proponents of marginal likelihood mainly counter cross-validation by focusing on its inability to identify the "true" model. For example, Gronau and Wagenmakers (2019) used Beta-Bernoulli models to generate simulated data in experiments, fitted data with models of varying complexity, and evaluated and compared models using LOO-CV and Pseudo-Bayes factors calculated from LOO-CV. Results showed that besides LOO-CV's inherent flaw of selecting more complex models rather than the true model that generated the data, LOO-CV's support for the true model shows an inverted U-shape as data increase. As data grow, LOO-CV's support for the true model first decreases then increases. Therefore, Gronau and Wagenmakers argue that researchers should be extremely cautious when using LOO-CV.

Vehtari et al. (2019) refuted Gronau and Wagenmakers' (2019) view, arguing that M-Closed settings only simplify modeling problems and rarely occur in practical applications. Furthermore, Vehtari et al. (2019) believe Gronau and Wagenmakers incorrectly used LOO-CV to calculate Pseudo-Bayes factors. Conversely, if using stacking methods with LOO-CV from each model as input values, the calculated model weights can effectively select the optimal model in M-Closed environments.

On the other hand, proponents of cross-validation argue that although marginal likelihood has many excellent theoretical properties, its practical application is often unsatisfactory. The reason is that marginal likelihood is not a measure of model generalization ability but rather measures the model's ability to explain current data given prior distributions and the model. Even if a model uses appropriate prior distributions and has better marginal likelihood, its generalization ability on out-of-sample data may not be stronger than other models (Lotfi et al., 2022).

Furthermore, selecting appropriate prior distributions in Bayesian inference is extremely difficult. For example, Gelman, Carlin, et al. (2013) argue that in practical applications of marginal likelihood, inappropriate informative prior distributions can have enormous impact on marginal likelihood. The more non-informative the model's prior distribution, the more marginal likelihood tends to favor simpler models. In contrast, LOO-CV is not affected by this issue (Gel-

man, Carlin, et al., 2013). For example, Kennedy et al. (2019) tested the impact of different prior distributions on Bayes factors by modeling Balloon Analog Risk Task (BART) experimental data. They found that as prior distributions become more non-informative, Bayes factors gradually favor simpler models. This is also the case in our case study: because BIC assumes non-informative priors while Laplace approximation and Bridge Sampling use actual priors from fitted models, BIC's Bayes factor is much smaller than the other two.

6.2 Recommendations for Using Model Selection Indices

First, when conducting model comparison, we should pay attention to the applicable conditions for each index. Each model comparison index is only applicable in scenarios consistent with the modeling data. For example, AIC for DDM based on reaction time and choice data cannot be compared with AIC for reinforcement learning models based on choice data (Fontanesi et al., 2019).

Second, when relative indices cannot distinguish between different models, posterior predictive checks can also serve as a method for model selection. For example, Steingroever et al. (2014) found that in the Iowa Gambling Task, indices such as BIC had difficulty distinguishing between different models, while posterior predictive checks could effectively select the optimal model.

For instance, AIC and BIC, as the most common indices, are suitable for models with point estimation methods such as maximum likelihood estimation. However, the choice between AIC and BIC remains controversial.

BIC's penalty term is stronger, and as in our case study, it tends to select simpler models. Therefore, researchers can choose indices based on their research hypotheses about effect size and statistical power. For example, both Type I and Type II errors for BIC decrease as sample size increases. For AIC, Type II error decreases with sample size, but Type I error does not. Moreover, AIC's Type II error is smaller than BIC's (Dziak et al., 2020). That is, under equal sample conditions, AIC can identify models with better out-of-sample predictive ability as optimal but also risks larger Type I error. While BIC has the ability to identify true models, it also has higher Type II error—the probability of selecting a poorly performing model.

Using model recovery methods to decide which index to use is also an option (Wilson & Collins, 2019). For example, Collins and Frank (2012) used more complex models to simulate data and fitted the data with both complex and simple models. They found that when using BIC as the model comparison index, fitting results supported the simple model—that is, BIC often over-penalized complex models, failing to recover the true model underlying the simulated data, while AIC could recover the more complex true model (Collins & Frank, 2018). Finally, many researchers recommend reporting both AIC and BIC. If results from both are consistent, model comparison results are more reliable. If they contradict, discussions can be categorized based on different principles (Farrell & Lewandowsky, 2018).

Additionally, different parameter estimation methods also limit the use of model comparison methods. For models using Bayesian parameter estimation, we can use MCMC samples to calculate more precise approximation indices such as marginal likelihood or LOO-CV. If using point-estimate MAP, we can also use Laplace approximation to calculate marginal likelihood. With informative priors, marginal likelihood performs better than approximations to cross-validation such as WAIC. Evans (2019) used LBA models to compare the impact of different levels of informativeness in prior distributions on model comparison, finding that when prior distributions are non-informative or weakly informative, marginal likelihood tends to over-penalize complex models, causing results to deviate from optimal choices; when prior distributions are moderately informative, marginal likelihood results are closer to optimal and better than WAIC; when prior distributions are strongly informative, marginal likelihood tends to select overly complex models. Therefore, when we have sufficient knowledge about model priors and set informative priors, marginal likelihood may be a better choice; when using non-informative priors, or setting informative priors without certainty about their appropriateness, indices insensitive to priors such as WAIC, DIC, and LOO-CV are more appropriate.

6.3 New Developments in Model Comparison

The common way of using model comparison indices is to compare their sum or mean across all subjects. However, this approach ignores differences between subjects and the potential impact of extreme values on model comparison. Random Effects Bayesian Model Selection (RE-BMS), originating from Dynamic Causal Modelling (DCM) (Stephan et al., 2009), can effectively reduce the impact of extreme values and has been widely applied in cognitive modeling. RE-BMS uses Bayesian hierarchical models to consider subject differences, employing multinomial and Dirichlet distributions to avoid impacts from asymmetric data distributions. Additionally, RE-BMS introduces Protected Exceedance Probability (PXP), representing the probability that a given model's marginal likelihood is greater than or equal to other models and can serve as the "true model" generating current data—that is, $PXP = p(r_{m=j} \geq r_{m\neq j}|y)$. PXP greater than 0.95 can be considered, like traditional hypothesis testing, as indicating the model is significantly better than others (Iglesias et al., 2013). Notably, toolboxes such as SPM and VBA in Matlab and the `bmsR` package in R can all calculate PXP (Daunizeau et al., 2014), making it widely applicable in cognitive modeling. Additionally, when using information criterion indices such as AIC and BIC as inputs for RE-BMS, these indices need to be divided by -2 to ensure correct results.

References

胡传鹏, 孔祥祯, 彭凯平. (2018). 贝叶斯因子及其在 JASP 中的实现. 心理科学进展, 26(6), 951-965. https://doi.org/10.3724/sp.J.1042.2018.00951

区健新, 吴寅., 刘金婷, 李红. (2020). 计算精神病学：抑郁症研究和临床应用的新视角. 心理科

学进展, 28(1), 111-127. https://doi.org/10.3724/sp.J.1042.2020.00111

王允宏, van den Berg, D., Aust, F., Ly, A., Wagenmaker, E.-J., 胡传鹏. (2023). 贝叶斯方差分析在 JASP 中的实现. 心理技术与应用, 11(9), 528-541. http://dx.doi.org/10.16842/j.cnki.issn2095-5588.2023.09.002

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., & Martin, O. A. (2023). PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, e1516. https://doi.org/10.7717/peerj-cs.1516

Acerbi, L., Dokka, K., Angelaki, D. E., & Ma, W. J. (2018). Bayesian comparison of explicit and implicit causal inference strategies in multisensory heading perception. *PLoS Computational Biology*, 14(7), e1006110. https://doi.org/10.1371/journal.pcbi.1006110

Ahn, W. Y., Haines, N., & Zhang, L. (2017). Revealing Neurocomputational Mechanisms of Reinforcement Learning and Decision-Making With the hBayesDM Package. *Computational Psychiatry*, 1, 24-57. https://doi.org/10.1162/CPSY_a_{00002}

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Allwein, E. L., Schapire, R. E., & Singer, Y. (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1(2), 113-141. https://doi.org/10.1162/15324430152733133

Anderson, D., & Burnham, K. (2004). *Model selection and multi-model inference* (Vol. 63). Second. NY: Springer-Verlag.

Ballard, I. C., Wagner, A. D., & McClure, S. M. (2019). Hippocampal pattern separation supports reinforcement learning. *Nature Communications*, 10(1), 1073. https://doi.org/10.1038/s41467-019-08959-1

Betts, M. J., Richter, A., de Boer, L., Tegelbeckers, J., Perosa, V., Baumann, V., Chowdhury, R., Dolan, R. J., Seidenbecher, C., Schott, B. H., Duzel, E., Guitart-Masip, M., & Krauel, K. (2020). Learning in anticipation of reward and punishment: perspectives across the human lifespan. *Neurobiology of Aging*, 96, 49-57. https://doi.org/10.1016/j.neurobiolaging.2020.08.011

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518), 859-877. https://doi.org/10.1080/01621459.2017.1285773

Boehm, U., Evans, N. J., Gronau, Q. F., Matzke, D., Wagenmakers, E.-J., & Heathcote, A. J. (2023). Inclusion Bayes factors for mixed hierarchical diffusion decision models. *Psychological Methods*, Advance online publication. https://doi.org/10.1037/met0000582

Boehm, U., Annis, J., Frank, M. J., Hawkins, G. E., Heathcote, A., Kellen, D., Krypotos, A.-M., Lerche, V., Logan, G. D., Palmeri, T. J., van Ravenzwaaij, D., Servant, M., Singmann, H., Starns, J. J., Voss, A., Wiecki, T. V., Matzke, D., & Wagenmakers, E.-J. (2018). Estimating across-trial variability parameters of the Diffusion Decision Model: Expert advice and recommendations. *Journal of Mathematical Psychology*, 87, 46-75. https://doi.org/10.1016/j.jmp.2018.09.004

Bos, C. S. (2002). A comparison of marginal likelihood computation methods. *Compstat: Proceedings in Computational Statistics*.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. *Cognitive psychology*, 57(3), 153-178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261-304. https://doi.org/10.1177/0049124104268644

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 1-32. https://doi.org/10.18637/jss.v076.i01

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Cengage Learning.

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal theta overrides pavlovian learning biases. *Journal of Neuroscience*, 33(19), 8541-8548. https://doi.org/10.1523/JNEUROSCI.5754-12.2013

Clyde, M. A., Ghosh, J., & Littman, M. L. (2011). Bayesian Adaptive Sampling for Variable Selection and Model Averaging. *Journal of Computational and Graphical Statistics*, 20(1), 80-101. https://doi.org/10.1198/jcgs.2010.09049

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, 35(7), 1024-1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

Collins, A. G. E., & Frank, M. J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, 115(10), 2502-2507. https://doi.org/10.1073/pnas.1720963115

Daniel, R., Radulescu, A., & Niv, Y. (2020). Intact Reinforcement Learning But Impaired Attentional Control During Multidimensional Probabilistic Learning in Older Adults. *The Journal of Neuroscience*, 40(5), 1084-1096. https://doi.org/10.1523/JNEUROSCI.0254-19.2019

Daunizeau, J., Adam, V., & Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. *PLoS Computational Biology*, 10(1), e1003441. https://doi.org/10.1371/journal.pcbi.1003441

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, Pennsylvania, USA. https://doi.org/10.1145/1143844.1143874

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision making, affect, learning: Attention performance XXIII* (Vol. 23).

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, 19(8), 1153-1160. https://doi.org/10.1016/j.neunet.2006.03.002

Devine, S., Falk, C. F., & Fujimoto, K. A. (2023). Comparing the Accuracy of Three Predictive Information Criteria for Bayesian Linear Multilevel Model Selection. *PsyArXiv*. https://doi.org/10.31234/osf.io/p2n8a

Dickey, J. (1973). Scientific Reporting and Personal Probabilities: Student's Hypothesis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 35(2), 285-305. https://doi.org/10.1111/j.2517-6161.1973.tb00959.x

Dickey, J. M. (1976). Approximate posterior distributions. *Journal of the American Statistical Association*, 71(355), 680-689. https://doi.org/10.2307/2285601

Donkin, C., Heathcote, A., & Brown, S. (2009). Is the linear ballistic accumulator model really the simplest model of choice response times: A Bayesian model complexity analysis. *Ninth International Conference on Cognitive Modeling—ICCM2009*, Manchester.

Dorfman, H. M., & Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature Communications*, 10(1), 1-11. https://doi.org/10.1038/s41467-019-13737-7

Doucet, A., & Johansen, A. M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering*, 12(656-704), 3.

Dziak, J. J., Coffman, D. L., Lanza, S. T., Li, R., & Jermiin, L. S. (2020). Sensitivity and specificity of information criteria. *Briefings in Bioinformatics*, 21(2), 399-412. https://doi.org/10.1093/bib/bbz016

Evans, N. J. (2019). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychonomic Bulletin & Review*, 26(4), 1070-1098. https://doi.org/10.3758/s13423-018-01563-9

Evans, N. J., Hawkins, G. E., & Brown, S. D. (2020). The role of passing time in decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 316-326. https://doi.org/10.1037/xlm0000725

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2), 489-496. https://doi.org/10.1093/biomet/asaa028

Fontanesi, L., Gluth, S., Spektor, M. S., & Rieskamp, J. (2019). A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic Bulletin & Review*, 26(4), 1099-1121. https://doi.org/10.3758/s13423-018-1554-2

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, 67, 641-666. https://doi.org/10.1146/annurev-psych-122414-033645

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer series in statistics New York.

Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100(1), 70-87. https://doi.org/10.1016/j.jphysparis.2006.10.001

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage*, 34(1), 220-234. https://doi.org/10.1016/j.neuroimage.2006.08.035

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365), 153-160. https://doi.org/10.1080/01621459.1979.10481632

Gelfand, A. E., & Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3), 501-514. https://doi.org/10.1111/j.2517-6161.1994.tb01996.x

Gelman, A., Hwang, J., & Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997-1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC.

Geng, H., Chen, J., Chuan-Peng, H., Jin, J., Chan, R. C. K., Li, Y., Hu, X., Zhang, R.-Y., & Zhang, L. (2022). Promoting computational psychiatry in China. *Nature Human Behaviour*, 6(5), 615-617. https://doi.org/10.1038/s41562-022-01328-4

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, 71, 1-6. https://doi.org/10.1016/j.jmp.2016.01.006

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian Leave-One-Out Cross-Validation for model selection. *Computational Brain & Behavior*, 2(1), 1-11. https://doi.org/10.1007/s42113-018-0011-7

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E. J., & Steingroever, H. (2017). A

tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81, 80-97. https://doi.org/10.1016/j.jmp.2017.09.005

Guitart-Masip, M., Huys, Q. J., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: interactions between affect and effect. *Neuroimage*, 62(1), 154-166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Prentice Hall.

Hammersley, J. (2013). *Monte carlo methods*. Springer Science & Business Media.

Heck, D. W. (2019). A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72(2), 316-333. https://doi.org/10.1111/bmsp.12150

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3(2), 200-215. https://doi.org/10.1177/2515245919898657

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2), 297-307. https://doi.org/10.1093/biomet/76.2.297

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, 19(3), 404-413. https://doi.org/10.1038/nn.4238

Huys, Q. J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Computational Biology*, 7(4), e1002028. https://doi.org/10.1371/journal.pcbi.1002028

Iglesias, S., Mathys, C., Brodersen, K. H., Kasper, L., Piccirelli, M., den Ouden, H. E., & Stephan, K. E. (2013). Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron*, 80(2), 519-530. https://doi.org/10.1016/j.neuron.2013.09.009

Ikink, I., Engelmann, J. B., van den Bos, W., Roelofs, K., & Figner, B. (2019). Time ambiguity during intertemporal decision-making is aversive, impacting choice and neural value coding. *Neuroimage*, 185, 236-244. https://doi.org/10.1016/j.neuroimage.2018.10.008

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773-795. https://doi.org/10.1080/01621459.1995.10476572

Kennedy, L., Simpson, D., & Gelman, A. (2019). The Experiment is just as Important as the Likelihood in Understanding the Prior: a Cautionary Note on Robust Cognitive Modeling. *Computational Brain & Behavior*, 2(3-4), 210-217. https://doi.org/10.1007/s42113-019-00051-0

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7), 319-326. https://doi.org/10.1016/j.tics.2006.05.003

Kvålseth, T. O. (1985). Cautionary Note about R^2. *The American Statistician*, 39(4), 279-285. https://doi.org/10.1080/00031305.1985.10479448

Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLoS Computational Biology*, 15(4), e1006973. https://doi.org/10.1371/journal.pcbi.1006973

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, 14(10), 1250-1252. https://doi.org/10.1038/nn.2904

Li, J. A., Dong, D., Wei, Z., Liu, Y., Pan, Y., Nori, F., & Zhang, X. (2020). Quantum reinforcement learning during human decision-making. *Nature Human Behaviour*, 4(3), 294-307. https://doi.org/10.1038/s41562-019-0804-2

Li, Z.-W., & Ma, W. J. (2021). An uncertainty-based model of the effects of fixation on choice. *PLOS Computational Biology*, 17(8), e1009190. https://doi.org/10.1371/journal.pcbi.1009190

Lotfi, S., Izmailov, P., Benton, G., Goldblum, M., & Wilson, A. G. (2022). Bayesian Model Selection, the Marginal Likelihood, and Generalization. *Proceedings of the 39th International Conference on Machine Learning*. https://proceedings.mlr.press/v162/lotfi22a.html

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

McFadden, D. L. (1984). Chapter 24 Econometric analysis of qualitative response models. In *Handbook of Econometrics* (Vol. 2, pp. 1395-1457). Elsevier. https://doi.org/10.1016/S1573-4412(84)02016-3

Menard, S. (2000). Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician*, 54(1), 17-24. https://doi.org/10.1080/00031305.2000.10474502

Meng, X.-L., & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 831-860. https://www.jstor.org/stable/24306045

Merlise, C., & Edward, I. G. (2004). Model Uncertainty. *Statistical Science*, 19(1), 81-94. https://doi.org/10.1214/088342304000000035

Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. *Trends in Cognitive Sciences*, 16(1), 72-80. https://doi.org/10.1016/j.tics.2011.11.018

Murphy, K. P. (2023). *Probabilistic machine learning: an introduction*. MIT press.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1), 79-95. https://doi.org/10.3758/BF03210778

Myung, J., & Pitt, M. (2018). Model comparison in psychology. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Vol. 5, pp. 1-34). https://doi.org/10.1002/9781119170174.epcn503

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, 21(6), 425-433. https://doi.org/10.1016/j.tics.2017.03.011

Pedersen, M. L., Ironside, M., Amemori, K. I., McGrath, C. L., Kang, M. S., Graybiel, A. M., Pizzagalli, D. A., & Frank, M. J. (2021). Computational phenotyping of brain-behavior dynamics underlying approach-avoidance conflict in major depressive disorder. *PLoS Computational Biology*, 17(5), e1008955. https://doi.org/10.1371/journal.pcbi.1008955

Plummer, M., Stukalov, A., & Denwood, M. (2016). rjags: Bayesian graphical models using MCMC. *R package version*, 4(6).

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4), 260-281. https://doi.org/10.1016/j.tics.2016.01.007

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306), 1593. https://doi.org/10.1126/science.275.5306.1593

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461-464. https://www.jstor.org/stable/2958889

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3), 333-343. https://doi.org/10.1007/BF02294360

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2020). Uncertainty in Bayesian leave-one-out cross-validation based model comparison. *arXiv*. https://doi.org/10.48550/arXiv.2001.00980

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 485-493. http://www.jstor.org/stable/24774528

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, 16(7), 966-973. https://doi.org/10.1038/nn.3413

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models of the Iowa Gambling Task. *Decision*, 1(3), 161-183. https://doi.org/10.1037/dec0000005

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2016). Bayes factors for reinforcement-learning models of the Iowa gambling task. *Decision*, 3(2), 115. https://doi.org/10.1037/dec0000040

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, 46(4), 1004-1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society: Series B*, 39(1), 44-47. https://doi.org/10.1111/j.2517-6161.1977.tb01603.x

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections. *Communications in Statistics-theory and Methods*, 7(1), 13-26. https://doi.org/10.1080/03610927808827599

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., Cheng, K., & Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron*, 74(6), 1125-1137. https://doi.org/10.1016/j.neuron.2012.04.030

Swart, J. C., Frobose, M. I., Cook, J. L., Geurts, D. E., Frank, M. J., Cools, R., & den Ouden, H. E. (2017). Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *Elife*, 6. https://doi.org/10.7554/eLife.22169

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., & Willemsen, J. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1), 1-26. https://doi.org/10.1038/s43586-021-00017-2

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological Methods*, 16(1), 44-62. https://doi.org/10.1037/a0021765

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In *The Oxford handbook of computational and mathematical psychology* (pp. 300-319). Oxford University Press.

Vehtari, A. (2022). Cross-validation FAQ. https://avehtari.github.io/modelselection/CV-FAQ.html

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413-1432. https://doi.org/10.1007/s11222-016-9696-4

Vehtari, A., Simpson, D. P., Yao, Y., & Gelman, A. (2019). Limitations of "Limitations of Bayesian Leave-one-out Cross-Validation for Model Selection". *Computational Brain & Behavior*, 2(1), 22-27. https://doi.org/10.1007/s42113-018-0020-6

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., & Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1), 3581-3618. http://jmlr.org/papers/v17/14-540.html

Verstynen, T., & Kording, K. P. (2023). Overfitting to 'predict' suicidal ideation. *Nature Human Behaviour*, 7(5), 680-681. https://doi.org/10.1038/s41562-023-01560-6

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17(2), 228-243. https://doi.org/10.1037/xlm0000725

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779-804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192-196. https://doi.org/10.3758/BF03206482

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, 60(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(12). http://jmlr.org/papers/v11/watanabe10a.html

Westbrook, A., van den Bosch, R., Määttä, J., Hofmans, L., Papadopetraki, D., Cools, R., & Frank, M. J. (2020). Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science*, 367(6484), 1362-1366. https://doi.org/10.1126/science.aaz5891

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for

testing composite hypotheses. *The annals of mathematical statistics*, 9(1), 60-62. http://www.jstor.org/stable/2957648

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, 8. https://doi.org/10.7554/eLife.49547

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 92(4), 937-950. https://doi.org/10.1093/biomet/92.4.937

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3), 917-1007. https://doi.org/10.1214/17-BA1091

Zhang, L., Lengersdorff, L., Mikus, N., Glascher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: frameworks, pitfalls and suggestions of best practices. *Social Cognitive and Affective Neuroscience*, 15(6), 695-707. https://doi.org/10.1093/scan/nsaa089

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1), 95-112. https://doi.org/10.1016/j.jeconom.2015.02.006

Note: Figure translations are in progress. See original paper for figures.