nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02348-6

Addressing low statistical power in computational modelling studies in psychology and neuroscience

Received: 28 August 2024

Accepted: 10 October 2025

Published online: 17 November 2025

Check for updates

Payam Piray

Computational modelling is a powerful tool for uncovering hidden processes in observed data, yet it faces underappreciated challenges. Among these, determining appropriate sample sizes for computational studies remains a critical but overlooked issue, particularly for model selection analyses. Here we introduce a power analysis framework for Bayesian model selection, a method widely used to choose the best model among alternatives. Our framework reveals that while power increases with sample size, it decreases as more models are considered. Using this framework, we empirically demonstrate that psychology and human neuroscience studies often suffer from low statistical power in model selection. A total of 41 of 52 studies reviewed had less than 80% probability of correctly identifying the true model. The field also heavily relies on fixed effects model selection, which we demonstrate has serious statistical issues, including high false positive rates and pronounced sensitivity to outliers.

In recent years, advances in computational methods combined with the availability of online platforms have fuelled a substantial increase in computational research within the behavioural sciences, particularly in cognitive science, psychiatry and neuroscience[1–9]. This shift has been transformative, elevating computational modelling from a niche approach to a common tool for uncovering the hidden processes that drive behavioural and neural data[10,11]. By creating sophisticated computational models that simulate cognitive processes, researchers can test their hypotheses and gain insights into the underlying mechanisms of human behaviour and brain function. The impact of this computational revolution extends far beyond mere data analysis. It has fundamentally altered the way scientists conceptualize and investigate complex cognitive phenomena. For instance, in decision-making research, computational models have revealed how the brain weighs multiple sources of information to arrive at choices, shedding light on both normal cognitive function and its aberrations in conditions like addiction or anxiety disorders[12–15]. In neuroscience, these models have helped bridge the gap between cellular-level processes and large-scale brain dynamics, offering a more integrated understanding of neural systems[16–18]. However, these advancements bring new challenges[19]. Researchers must grapple with issues of reproducibility, generalizability, and robustness in computational studies. As the field moves forward, addressing these challenges will be crucial to ensuring that computational models continue to drive meaningful progress in our understanding of the human mind and brain.

One common approach for statistical inference in computational studies is to present evidence for a specific computational model by demonstrating that the model of interest provides a better fit to experimental data compared with alternative models. This is achieved through a statistical inference method known as Bayesian model selection[20]. For example, a researcher studying decision-making might compare different types of reinforcement learning model to determine which one better explains participants' choices in a task[21]. Model selection is crucial because it allows researchers to evaluate the relative merits of different computational theories and select the one that best accounts for the observed data. It is a general tool that has been widely proposed as a replacement for classical null hypothesis testing[22–26] and lies at the heart of practices required for making inferences about data using computational models[27–31]. Importantly, researchers often use Bayesian model selection techniques to compare several models simultaneously.

Department of Psychology, University of Southern California, Los Angeles, CA, USA. e-mail: piray@usc.edu

Intuitively, the accuracy of model selection depends not only on the amount of data available (the sample size) but also on the number of competing explanations being considered (the number of candidate models). To illustrate this concept, consider trying to determine a country's favourite food. If the country's culture is less food-oriented, such as the Netherlands, you may only need to choose between a few options, say 'stamppot' or 'erwtensoep'. In this case, it would not take much information to confidently identify one as more favoured than the other. You could ask a relatively small sample of people, and their responses would probably provide a clear probability distribution favouring one dish. However, if your task is to detect the favourite food in Italy, you face a much more complex challenge due to the dozens of candidate dishes. With so many plausible options, each with different characteristics, it becomes increasingly difficult to single out the best one with a limited sample size. To reduce the uncertainty caused by having numerous candidates, you would need a substantially larger sample size to reliably identify the most favoured dish. Similarly, in scientific research, when there are many plausible models involved, each making different predictions about the phenomenon of interest, it becomes increasingly difficult to confidently select the best model with limited data. The more candidate models under consideration, the larger the typical sample size required to distinguish among them accurately.

In this Article, we argue that low statistical power for model selection is a major yet under-recognized issue in computational modelling research within the behavioural sciences, particularly in psychology and neuroscience, even for studies with large sample sizes. Despite its prevalence, there is a concerning lack of awareness about crucial factors influencing statistical power for model selection. This is important because, similar to any other statistical inference, model selection with low statistical power has a reduced chance of detecting true effects (type II errors)[32]. Moreover, although largely underappreciated, underpowered studies also reduce the likelihood that a statistically significant finding reflects a true effect (type I error)[33,34]. The main reason for this power deficiency appears to be that researchers often fail to account for how expanding the model space reduces power for model selection.

Another critical issue in model selection for computational studies, which may also be a contributing factor in the power deficiency problem, is the prevalence of the classical 'fixed effects' approach to model selection, which neglects between-subject variability in model expression[35–40]. Although this method often yields a clear winning model, it has been deemed as implausible in the neuroimaging community over the past decade[37,41]. This is because the fixed effects approach makes the a priori assumption that a single model is the true underlying model for all subjects in a study. This disregards between-subject variability in model validity, relying on the strong (and potentially unwarranted) assumption of uniformity across the population. Despite this conceptual deficiency, fixed effects model selection remains ubiquitous in psychological studies, especially in cognitive science. Here, we argue that there are more worrying practical issues that render this approach virtually always incorrect: its lack of specificity resulting in unreasonably high rates of false positives and its extreme sensitivity to outliers.

We first present a statistical framework for power analysis in model selection studies, which demonstrates that while statistical power increases with sample size, it decreases as the model space expands. This framework provides researchers with a general-purpose tool to calculate power before conducting their studies, given the size of their model space. Using this framework, we perform a narrative review of the literature, revealing that the field suffers from critically low statistical power for model selection. Our review also uncovers the widespread use of fixed effects model selection in the field. Thus, we analyse the issues of fixed effects model selection, both its high rates of false positives and its severe sensitivity to outliers[35,39,42] and argue that the field should avoid this approach in favour of more reliable random effects methods.

Results

Bayesian model selection

Consider a situation in which we have measured data in N participants, and we consider K alternative models as plausible models. Specifically, if $\mathbf{X}_n$ is the dataset for nth participant, we assume that we are able to obtain model evidence for each model k, $\ell_{nk} = p(\mathbf{X}_n|M_k)$ by marginalizing over parameters of the model. Model evidence gives a measure of goodness of fit that is properly penalized for model complexity, as the effects of model parameters are marginalized. In practice, one often needs to approximate model evidence as marginalizing exactly is usually infeasible[36]. However, efficient methods have been developed for approximating model evidence, from simple classic measures such as the Akaike Information Criterion[43] and Bayesian Information Criterion[44] or their modern counterparts[45] to more advanced techniques such as variational Bayes[46]. For our purposes here, we simply assume that approximate or exact model evidence is available for each subject and model.

In psychological sciences, we typically aim to make inferences about a population by testing a representative sample. Consequently, computational modellers seek to quantify evidence for each model across the entire sampled dataset. This process involves two potential assumptions about data generation at the population level[35,36]. The first assumption posits that only one model is expressed in the population, implying that any between-subject variability in model evidence is solely due to measurement noise. Although this assumption may be valid when there is no variability across subjects in the specific measured aspect, it is often a strong assumption for behavioural or neural data across a group of participants. This approach is reminiscent of fixed effects methods in general linear modelling, where data were concatenated across participant groups. Given the typical between-subject variability in psychological sciences, such an approach to general linear modelling is now considered inappropriate in modern psychology and neuroscience, despite having been the dominant method in early neuroimaging studies[47–49]. It is easy to see that the fixed effects model evidence across the group is given by the sum of log model evidence across all subjects

$$L_k = \sum_n \log \ell_{nk}, \tag{1}$$

where L_k is the (log) model evidence for model k. We will delve into specific problems of the fixed effects approach later, but for now, we focus on another widespread approach to model selection.

Random effects Bayesian model selection

The second method, called random effects model selection, accounts for variability across individuals in terms of which model best explains their behaviour[36–40]. This approach permits the possibility that different individuals may be best described by different models, and a key goal of the modeller is to quantify this heterogeneity. Statistically, unlike the fixed effects approach, the random effects assumption implies that between-subject variability stems not only from measurement noise but also from meaningful individual differences in the measured aspect. Thus, this method acknowledges the inherent variability in human populations, permitting a more nuanced understanding of cognitive processes or neural mechanisms.

Formally, random effects model selection estimates the probability that each model in a set of models is expressed across the population[36]. Consider a model selection problem with a model space of size K and sample size of N. We define a random variable, $\mathbf{m}$, a 1-by-K vector where each element m_k represents the probability that model k is expressed in the population. The goal of the modeller is to estimate this probability or, at the very least, ensure that it is significantly higher for the model of interest.

We assume $\mathbf{m}$ follows a Dirichlet distribution $p(\mathbf{m}) = \mathrm{Dir}(\mathbf{m}|\mathbf{c})$, where $\mathbf{c}$ is a 1-by-K vector with all elements set to $c = 1$ (a standard choice

https://doi.org/10.1038/s41562-025-02348-6

that assumes equal prior probability for all models, consistent with the literature[36–39]). The experimental group sample is then generated based on **m** and N according to a multinomial distribution. This means that each participant's data are generated independently by exactly one of the models, with the probability of model k being expressed given by m_k. This approach differs from the fixed effects method, where one model was expressed across all subjects according to a categorical distribution. In the random effects approach, one model is expressed per subject, which across the group can be modelled using the multinomial distribution (essentially, this is equivalent to N independent categorical distributions multiplied together).

The goal of random effects model selection is to infer the posterior probability distribution over the model space **m**. Given model evidence values for all models and all participants, this posterior is given by a Dirichlet distribution

$$p(\mathbf{m}|\mathbf{D}) = \mathrm{Dir}(\mathbf{m}|\overline{\mathbf{N}} + \mathbf{c}), \tag{2}$$

where **D** denotes the full dataset (essentially model evidence ℓ_{nk} for all models and all participants)[36]. Here $\overline{\mathbf{N}}$ is a 1-by-K vector in which $\overline{N}_k$ indicates the effective number of subjects explained by model k and it is given by

$$\overline{N}_k = \sum_n^N r_{nk}, \tag{3}$$

$$r_{nk} = \frac{\ell_{nk}}{\sum_j \ell_{nj}}, \tag{4}$$

where each r_{nk} (ranging between 0 and 1) can be interpreted as the probability that model k generated the data for participant n. The inference procedure, therefore, involves summing the probabilities r_{nk} across participants for each model, which functions similar to a soft or weighted count of how many participants are best explained by that model.

The posterior Dirichlet distribution over **m** contains all the information needed to perform random effects model selection. One commonly used summary is the model frequency, defined as the expected value of each element, $\mathbb{E}[m_k]$. Another widely used metric is the exceedance probability, which quantifies the belief that a given model is more likely than all others[36–38]. Formally, it is defined as $\varphi_k = \Pr\left[m_k > m_j|\mathbf{D}\right]$ for all $j \neq k$, where $\Pr[\cdot]$ denotes probability. This probability thus represents how confident we can be that the corresponding model is the best across the population. Similar to model frequency, the exceedance probability solely depends on the parameters of the posterior Dirichlet distribution.

In this Article, we base our power analysis method on the random effects model selection framework. This choice reflects both conceptual and practical considerations: random effects assumptions better match how psychological datasets are generated and are less sensitive to outliers than fixed effects methods, as we will demonstrate. We begin by presenting our power analysis method formally.

Power analysis for model selection

Power is defined as the probability of detecting an effect if that effect truly exists. Thus, we simulate scenarios in which there is a truly best model in the population and assess how reliably this model can be recovered using random effects model selection. This approach provides a framework for power analysis in model selection that requires no prior data or specific assumptions about the models themselves.

Without loss of generality, we designate model 1 as the true model. Using a Monte Carlo approach, we draw samples from a Dirichlet distribution and retain only those where model 1 is the truly best model across the population (that is, $m_1^l > m_k^l$ for each sample l and all $k \neq 1$). For each retained population sample, we simulate a group sample by drawing N participants from a multinomial distribution based on the corresponding population distribution.

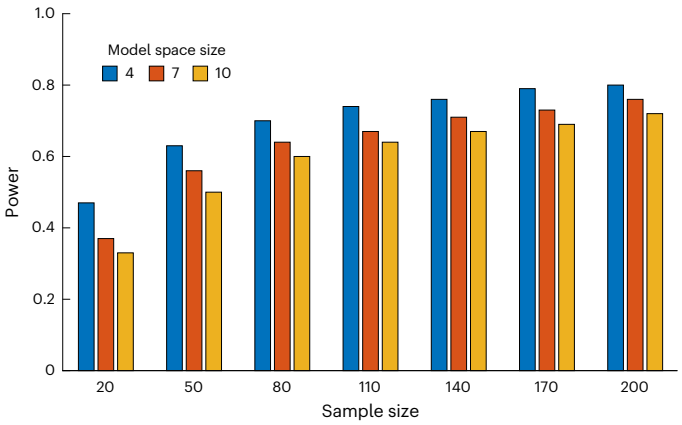


Fig. 1 | Power as a function of sample size and size of model space. The y axis represents power, calculated for various sample sizes (shown on the x axis). The size of the model space (four, seven or ten models) is depicted using different colours. The analysis shows that statistical power in model selection generally increases with larger sample sizes and decreases with a greater number of competing models.

Next, we perform Bayesian model selection and assess how reliably the true model can be recovered. We use exceedance probability as the primary metric because it quantifies how confident we can be that one model is better than all others in the population. This aligns naturally with how researchers think about model selection: we want to know the probability that the selected model is truly the best one.

For each group sample, we perform Bayesian inference (equations (2)–(4)) by computing the posterior Dirichlet parameters. We then compute the exceedance probability of model 1 given the simulated data, φ_1. Power is then calculated as the percentage of simulations where model 1 is declared the winner, that is, when its exceedance probability exceeds a critical threshold (that is, $\varphi_1 > \varphi_{\mathrm{crit}}$).

This procedure requires a threshold φ_{crit} to determine when evidence is 'significant' enough to declare a model as the winner. The choice of φ_{crit} directly affects the balance between sensitivity (controlling type II errors) and specificity (controlling type I errors): a lenient threshold will show inflated power by allowing too many false positives, while an overly strict threshold will reduce power by missing true effects. Previous implementations of random effects model selection have lacked clear guidelines for establishing this threshold. We address this gap by calibrating φ_{crit} to control false positive rates at 5%, paralleling standard practice in classical hypothesis testing.

To this end, we use a Monte Carlo simulation under the null hypothesis where all models are equally likely. Specifically, we simulate null scenarios with N participants and K models, generating group samples from a multinomial distribution with equal probabilities ($1/K$) for each model. For each simulation, we calculate the posterior exceedance probabilities for all models. We then set φ_{crit} as the threshold that yields a false positive rate of 5% or less, that is, in only 5% of null simulations does any model's exceedance probability exceed φ_{crit}. This calibrated threshold is then used in our power calculation above.

Simulation analyses

We tested our power analysis method through simulations that varied two key factors: sample size and the number of competing models. Specifically, we examined model spaces containing four, seven or ten models across a range of sample sizes. The results revealed two critical patterns (Fig. 1). First, as expected, power increased with sample size regardless of how many models were being compared. Second, and perhaps more striking, the number of models in the comparison affected

https://doi.org/10.1038/s41562-025-02348-6

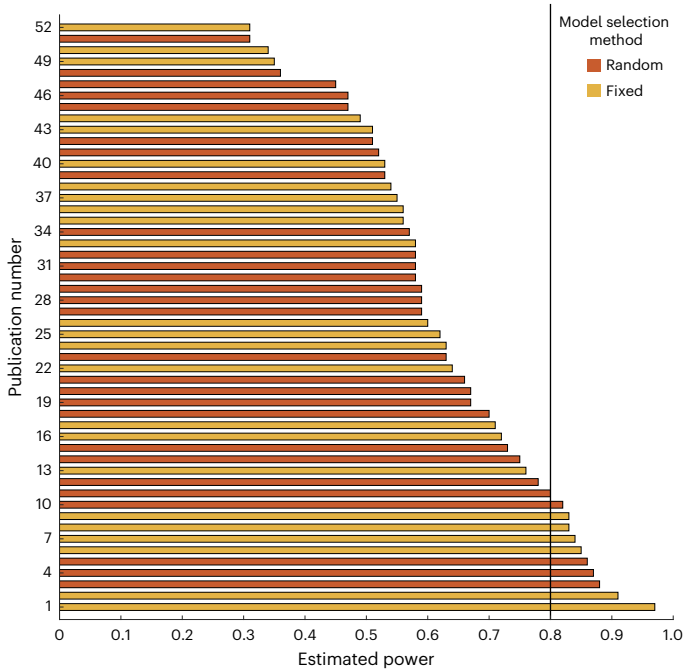


Fig. 2 | Narrative review of the literature. Estimated power for 52 reviewed studies based on their sample sizes and model space sizes. Among these, 41 studies fell below the standard 0.8 power threshold. The plot also indicates whether each study used random or fixed effects model selection.

power: adding more models to the model space reduced power even when sample size stayed the same. Overall, these analyses demonstrate how both sample size and model space size impact statistical power for model selection, with larger samples and smaller model spaces yielding greater power.

Narrative review of the literature

We used this framework to assess the current state of the fields of psychology and neuroscience with regard to power of model selection. We sought to assess the power in recent computational modelling studies published in high-profile journals (*Nature Human Behaviour*, *Nature* and *Science*) within the field of psychology and human neuroscience. We performed a narrative literature review of all behavioural and neural studies published in the last 5 years in these three journals. This yielded 54 studies through searches on the journal websites using permutations of query terms related to Bayesian model selection (Methods). Two studies with a sample size of only five were excluded from the analyses, as their small sample sizes probably reflect constraints on recruitment rather than typical study design practices. Our analysis of these studies' statistical power yielded concerning results: given their actual sample sizes and model spaces, 41 of the 52 studies (79%) had less than 80% probability of correctly identifying the true model, falling short of conventional standards (Fig. 2). In other words, these studies typically tested a much bigger model space than was supported by their sample size.

Our primary results demonstrate that power decreases as model space size increases. This pattern emerges because, when the probability space must be divided among more models, each model necessarily captures a smaller proportion, leading to reduced 'effect sizes.' For instance, when sampling from a uniform Dirichlet distribution and retaining only cases where one model dominates, the average effect size (the difference between the best and the second-best models in the population) is 0.33 for $K = 3$ models but drops to 0.25 for $K = 4$ models. This natural scaling reflects how model spaces typically evolve: researchers tend to add competitive models that compete for similar variance rather than irrelevant alternatives[50].

However, one might argue that this natural reduction in effect sizes is not inevitable in computational studies. To test whether our findings only reflect reduced effect sizes, we conducted additional analyses holding effect sizes constant regardless of model space size. We calibrated the Dirichlet parameter to generate true population scenarios with predetermined effect sizes, systematically adjusting the parameter until the mean difference between the best and second-best model matched our target effect size value. Under these conditions, power no longer decreased with model space size, confirming that the dependency of power on model space size operates through effect size scaling.

Importantly, even with constant effect sizes, our narrative review showed that 31 of 52 studies (60%) were underpowered for medium effects (Extended Data Fig. 1). The effect size was defined according to conventional standards for binomial tests[51]. These findings indicate that our main results are robust even when effect sizes are held constant. Although the constant effect size scenario may not reflect typical research practice, where adding models usually introduces competitive alternatives rather than weak distractors, our analyses suggest that the field would face substantial power challenges even under these more generous conditions.

It is also noteworthy that 46% of the reviewed studies (24 out of 52) used fixed effects model selection, which, as explained earlier, is highly sensitive to outliers. We classified studies as using fixed effects only if they effectively disregarded variability across individuals in terms of model evidence. Studies that employed any method appreciating between-subject variability in model selection, even if not using the more advanced random effects approach[36,39,40], were considered as using random effects model selection. We now demonstrate the severity of fixed effects' issues through simulation analyses where the ground truth is known.

Fixed effects model selection is statistically unreliable

Our power analysis approach and results are based on the assumptions of the random effects approach. An alternative strategy for model selection, which does not necessarily result in the same form of power, is the fixed effects approach that assumes a single model underlies data across all subjects. Although the fixed effects approach has faced criticism for failing to properly account for between- versus within-subject variability[35–39], one could still argue that modelling this source of variability is not essential for model selection, and the assumption of a common model across all subjects may be reasonable. Here, instead of focusing on whether generative assumptions of the fixed effects are valid, we focus on its two practical issues, both of which arise from a lack of proper treatment of between-subject variability. The first issue is that the fixed effects is prone to extremely high false positive rates. In other words, even when the correct model is not in the model space, and thus, any difference between models is due to noise, the fixed effects approach declares one of the models as the winner. The second issue is that the fixed effects model selection is highly susceptible to the influence of outliers[35,39,42].

To demonstrate these issues, we performed a prototypical simulation analysis. Consider a situation where $N = 50$, the model space consists of three models ($K = 3$), and for each participant, there are $T = 200$ data points. First, we assume that none of the models gives a good account of the data, or in other words, the true model is not included in the model space (this is reminiscent of simulation under the null in the language of classical statistics). Thus, for all subjects, the probability of each data point under all models, that is, the model evidence, was drawn randomly between 0 and 1. Trial-level data were then used to calculate log-evidence for each simulated participant and model, which is given by the sum of log probabilities. These were then subject to Bayesian model selection. For the fixed effects model selection, a model is usually considered the winner of model selection if its log-evidence is larger than the other models by a threshold of 3, which is

https://doi.org/10.1038/s41562-025-02348-6

Table 1 | The fixed effects model selection demonstrates an extremely high rate of false positives

	Model 1	Model 2	Model 3
True model	0	0	0
Fixed effects model selection	33%	32%	32%
Random effects model selection	1%	1%	1%

The values indicate the percentage of simulations in which each model was selected as the winner. Even though the simulation was conducted under the null hypothesis, where no model was more likely at the population level, fixed effects model selection declared one of the three models as the winner in 97% of simulations, with approximately equal probability for each model. By contrast, the random effects approach almost never declared any model as the winner.

Table 2 | Fixed effects model selection is sensitive to extreme outliers

		Model 1	Model 2	Model 3
Simulation 1	True model	0	0	0
	Winner with the outlier	85%	7%	7%
Simulation 2	True model	0	100%	0
	Winner with the outlier	77%	21%	0

The values indicate the percentage of simulations in which each model was selected as the winner. In simulation 1, there is no true model, while in simulation 2, model 2 is the true model across all simulations. In both simulations, one extreme outlier was designed to favour model 1.

roughly equal to a (group) Bayes factor of 20 that is considered strong evidence[30,35,37,38]. We repeated this process 10,000 times.

Ideally, since there was effectively no real difference between these models across synthetic subjects, we would expect none of the models to surpass the threshold for being declared the winner. However, the results were the opposite: in 97% of simulations, fixed effects model selection declared one of the three models as the winner. Given that there was no real difference between these models, each model was declared the winner roughly an equal number of times under fixed effects model selection. In other words, if you draw just one group from the population, the fixed effects approach will almost certainly identify one model as the winner—even though none of the models are more likely at the population level. If you repeat the experiment again, the fixed effects approach would declare again one of the models as the winner, although that model is most likely to be different from the previous time. This contrasts sharply with the random effects approach, which rarely declared a model as the winner in any of the simulations (Table 1).

Now, we examine the impact of outliers, that is, those who show strong evidence in favour of a model. Thus, we replace only one synthetic subject out of the 50 with an outlier in every simulation. We assume that the outlier shows a great match to one of the models, for example, model 1. Thus, for all subjects except the outlier, the probability of each data point under all models was generated randomly. However, for the outlier subject, the probability of data points under model 1 was very high—for example, randomly drawn between 0.9 and 1. We repeated this process 10,000 times to obtain the statistics of model 1 winning the model selection process. The simulation analysis revealed that model 1 won the model selection process in a staggering 84% of all simulations. This outcome, with model 1 being decisively favoured, can be attributed solely to the presence of just one outlier subject strongly favouring model 1 in the sample. Despite the probability of each data point for 49 subjects being simulated randomly for all three models, the fixed effects approach is so sensitive to outliers that a single outlier subject carried enough weight to dramatically sway model selection in favour of model 1 across the vast majority of simulated datasets. Again, this contrasts with the random effects approach, which almost never declared a model as the winner in the simulations.

In fact, the sensitivity of fixed effects model selection to outliers is so severe that outliers can flip the model selection results even when there is a clear underlying true model (simulation 2 in Table 2). To see this, we assume a slightly different scenario. Here, we assume that model 2 is the true underlying model for all subjects except for one outlier who shows a great deal of evidence for model 1, similar to the previous scenario. Therefore, we first conducted 10,000 simulations of 49 subjects in which model 2 was the fixed effects winner (with its log-evidence larger than the other models by a threshold of 3). Next, we simulated the outlier subject and added its evidence to the rest of the 49 subjects.

Ideally, we would expect that adding one outlier subject would not change the model selection results obtained based on the 49 subjects. However, the analysis revealed that model 1 still won the model selection process in 77% of cases with a threshold of 3. In other words, the presence of one outlier subject in a sample of 50 was sufficient to flip the results of the fixed effects model selection in 77% of simulations. This effect was not simply due to thresholding. Even when we raised the bar, considering a model the winner only if its log-evidence exceeded others by a threshold of 10, the results remained nearly unchanged. Model 1 still won in 73% of cases, driven by the presence of that single outlier subject.

One might argue that the extreme outliers simulated above were relatively rare in psychology and neuroscience. What was more likely in practice was an outlier that consistently showed modest evidence in favour of one particular model. To put this in numbers, we assumed that for the outlier subject, the probability of data points under model 1 was only slightly higher than for others, for example, ranging between 0.1 and 1.

We focused on the situation in which none of the models were true at the population level. Accordingly, for all other subjects, model probabilities were randomly drawn between 0 and 1 for all models. The simulation analysis revealed that, with only one modest outlier, the fixed effects approach declared model 1 the winner in about half of the simulations—much higher than when there was no outlier (Fig. 3). By adding one more outlier, the fixed effects approach declared model 1 as the winner in 64% of simulations. With five outliers (that is, only 10% of the sample showing modest evidence in favour of model 1), the fixed effects approach declared model 1 as the winner in 94% of simulations.

This scenario occurs quite often in practice: more complex models—often the ones of primary research interest—contain additional components that allow them to fit data from a small subset of subjects slightly better than chance. When using fixed effects model selection, these few subjects with slightly better fits could entirely determine the overall model selection results. This vulnerability arises because the fixed effects method, in contrast to the random effects approach, does not normalize evidence per subject but instead sums the log-evidence across all subjects (equation (1)).

While the random effects approach avoids these issues, it does have one consideration. The prior parameter c, if chosen unreasonably large, can result in overstating heterogeneity. However, with standard prior settings used here and in previous work[36,39,40] ($c = 1$), the random effects approach performs reliably well even if the assumptions of the fixed effects truly hold at the population level (see Supplementary Results for related simulation analyses).

Discussion

In this study, we developed a statistical framework for power analysis for model selection in computational studies. Our framework demonstrates that statistical power is not solely dependent on sample size but also on the size of the model space. Although increasing sample size leads to higher statistical power, expanding the model space has the opposite effect, reducing power. This finding has important implications for computational modelling research in the behavioural sciences,

https://doi.org/10.1038/s41562-025-02348-6

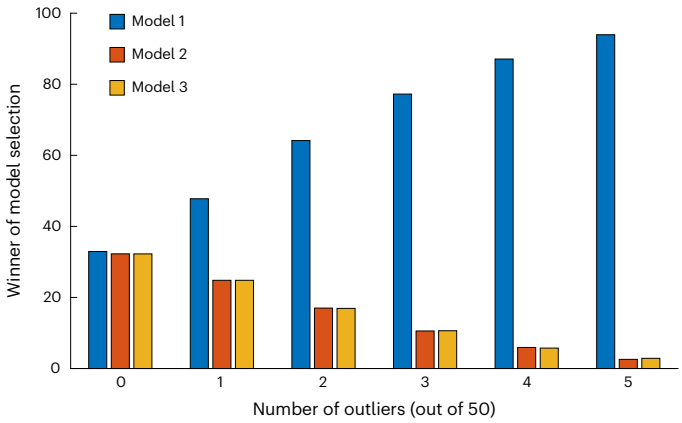


Fig. 3 | Fixed effects model selection is highly sensitive to modest outliers. The values on the *y* axis indicate the percentage of simulations in which each model was selected as the winner. Since none of the models are more likely than the others, the fixed effects approach selects each model approximately 33% of the time when there are no outliers, indicating a high rate of false positives. The inclusion of modest outlier subjects with a slight bias toward model 1 (where the probability of all trials for the outlier is between 0.1 and 1, while for others it is between 0 and 1) substantially biases model selection results toward model 1. With five outliers (that is, 10% of the group, as $N = 50$), the fixed effects model selection identifies model 1 as the winner in 94% of simulations.

as it suggests that studies with large sample sizes may still suffer from low statistical power if the model space is not carefully considered.

To address this issue, our framework provides researchers with a reliable tool to calculate power before conducting their study, enabling them to make informed decisions about sample size and model space. When we applied this framework to the current state of the field, we found that computational neuroscience and cognitive sciences suffer from a widespread lack of statistical power for model selection. This lack of statistical power can lead to both type I and type II errors[33,34], undermining the reliability and reproducibility of research findings.

Our analysis assumes model selection is the primary goal driving sample size decisions, which may not reflect actual research practice. In many studies, sample sizes are determined by other statistical tests central to the main hypotheses, with model selection serving as a secondary or exploratory analysis. In addition, researchers face practical constraints that our analysis does not consider, such as limited participant pools in patient populations and attrition in longitudinal studies. The model space itself often expands during the peer review as researchers add control models. Post hoc power analyses based on observed effects may also reveal adequate power for specific effects detected, which our a priori approach does not consider. Despite these limitations and recognizing that power is a probabilistic concept that does not deterministically validate or invalidate individual findings, our analyses nonetheless point to a systemic issue in the field that warrants attention.

To address this systemic issue, several strategies may help the field navigate power-related challenges within the practical constraints of research environments. Power analyses can be tailored specifically to model selection procedures, accounting for both sample size and the size of model space. When sample size is constrained by practical limitations, narrowing the model space to focus on the most theoretically relevant models may improve power. In addition, transparent reporting that distinguishes between model selection within the a priori set of models and post hoc model selection on models added after the main analyses can enhance the interpretability of findings. These approaches offer pathways to maximize the inferential value of model selection while acknowledging the real-world limitations researchers face.

A key methodological contribution of our framework is establishing principled thresholds for interpreting exceedance probabilities in random effects model selection. While exceedance probability has become a standard metric in computational modelling, the field has lacked consensus on what constitutes significant evidence. Previous work has generally treated exceedance probabilities like (one minus) *P* values but without proper calibration or consideration of false positive rates[37,39,40]. Our approach addresses this gap by introducing a decision threshold calibrated through null simulations to maintain an acceptable false positive rate (for example, 5%), paralleling classical hypothesis testing. This calibration is critical for power analysis because sensitivity (controlling type II errors) and specificity (controlling type I errors) trade off: lenient thresholds inflate power estimates while strict thresholds reduce them. By explicitly controlling type I errors, our method ensures that power estimates reflect genuine sensitivity rather than artifacts of arbitrary threshold choices. Our framework thus provides the field with both a practical tool for statistical power analysis and a principled approach to interpreting exceedance probabilities in studies using random effects Bayesian model selection.

The power analysis method outlined in this study is only based on the generative assumptions of the random effects account[36,40], making it a generic approach that can be used for any model selection study without requiring the specification of models. An alternative approach to random effects model selection is the fixed effects approach, which assumes that the data at the population level is generated by one of the models (unlike the random effects approach that assumes the data are generated at the individual level by one of the models in the model space). Although the fixed effects approach has been criticized for not properly accounting for within- and between-subject variability in model frequency[37,40], it is still widely used in the field. In this study, we highlight two important practical issues with this approach. First and foremost, our simulations show that fixed effects model selection frequently declares a winning model even when no true difference exists between models at the population level. In our simulations under the null hypothesis—where no model is truly superior—the fixed effects approach incorrectly identified one model as the winner in 97% of cases, with approximately equal probability across all models. This problematic scenario likely occurs frequently in practice. Consider a decision task where participants choose between two options on every trial. If the experimenter tests 50 participants, and none of the models in their candidate space truly captures the underlying processes, fixed effects analysis will almost certainly still declare one model winner. Should the researcher repeat the study with another 50 participants, fixed effects would again identify a winner, though probably a different one. In statistical terms, the fixed effects model selection approach (at least with conventional thresholds) has a false positive rate of 0.97, a rate that fundamentally undermines its reliability.

Second, as noted in the literature[35,42] and demonstrated here empirically, the fixed effects model selection is highly sensitive to outliers. The presence of even a single extreme outlier can dramatically influence the results of model selection when using fixed effects methods. Even presence of a few modest outliers, which slightly show more evidence towards one of the models, can completely drive the results of the fixed effects model selection. Note that this scenario happens quite often in practice: more complex models (often the ones of primary research interest) contain additional components that may allow them to fit the data from a small subset of subjects slightly better than chance. When using fixed effects model selection, these few subjects with slightly better fits can completely determine the overall model selection results.

Nevertheless, when comparing fixed and random effects approaches to model selection, it is important to understand the trade-offs. While random effects model selection is a slightly more complex representation of the data structure, which can make it prone to overstating heterogeneity, in practice, that is less concerning than the issues with the fixed effects that was highlighted here. First, in scenarios where one model truly dominates at the population level

(matching fixed effects assumptions), the random effects approach successfully identifies this model despite its theoretical flexibility. This occurs because the true effect size in such cases is substantial enough to be detected even with small sample sizes and large model spaces. Second, the random effects approach's performance is influenced by its prior parameter (c) for the Dirichlet distribution, particularly with smaller sample sizes. Our analyses show that with the standard setting of $c = 1$ commonly used in the literature, the random effects approach maintains appropriate sensitivity while still accommodating potential heterogeneity. As sample size increases, the influence of this prior diminishes. This balance allows the random effects approach to correctly identify dominant models when they exist while maintaining the flexibility to detect genuine heterogeneity when present, providing a more robust framework for model selection across diverse research contexts.

The method outlined in this study is a general approach that can be applied to any model selection study without requiring the specification of particular models. It estimates power based solely on sample size and the size of the model space, regardless of the nature of the data or the computational models under consideration. However, this generality also introduces certain limitations. Our power analysis assumes that each subject's data have been generated entirely by one of the models. In other words, while our approach assumes there might be between-subject variability in model expression (that is, different subjects might express different models), it assumes there is no within-subject variability in model expression (that is, all trials within a subject's dataset are generated based on the same model, and any within-subject variability is essentially due to noise). This approach is not optimal for situations in which models have different components that each explain some amount of within-subject variability. If these assumptions are violated, however—for example, when some trials are truly generated by one model and others by a different model—the presented method represents an optimistic scenario in which the estimated power serves as an upper bound. In most applications of Bayesian model selection, however, this limitation is not severe because the typical approach involves an increasing level of complexity in model space, with more complex models containing components that can account for the same variability that simpler models explain.

Relatedly, for even more accurate estimation of power, our framework can be augmented with a simulation-based approach that incorporates the specific models being considered, generating power estimates tailored to the particular model space. This method combines trial-wise simulations of models with our approach to drawing group-level samples from a population distribution. By simulating data under plausible parameter configurations, researchers can account for overlapping models and shared predictions, factors that may further reduce actual power relative to the best-case scenario. While this simulation-based method requires a priori assumptions about model parameters, which introduces its own uncertainties, it offers a valuable way to incorporate model-specific structure into power calculations. This combined approach enables researchers to tailor power analyses to their particular model space, more accurately estimating required sample sizes while preserving the general principles of our framework.

The core issue in Bayesian model selection is capturing variations in the frequency of different models across subjects. However, another key theoretical and practical challenge is capturing variations in the parameters of each model, as these parameters may also vary from subject to subject[39,52,53]. Model selection is often an intermediate step before thoroughly analysing the parameters of the winning or best-fitting model. This is particularly relevant in computational psychiatry, where it is common to test for potential differences between control and patient in specific parameters of the model of interest[54–57]. Future methodological work should provide a thorough and systematic analysis of this parameter variation issue. One potential approach is to augment the proposed power analysis method with systematic simulations across the model space. This would allow researchers to assess the sample size needed to achieve both accurate model selection and the desired effect size in terms of differences in model parameters between groups or conditions. It is important to note that while we focused here on the problem of model selection, there are other crucial steps to ensure the reproducibility of computational modelling studies, including model development, model recovery and model verification[11,31,58,59].

While this study focuses on model selection in computational modelling, the advantages of Bayesian model selection extend far beyond this specific domain. In fact, Bayesian model selection using Bayes factors is a powerful, general tool that has been widely proposed as a replacement for classical null hypothesis testing[22,24–26]. Bayes factors essentially represent the relative evidence for one model compared with another (for example, the model of interest versus the null model). However, Bayesian methods are equally valid and applicable for making inferences among more than two alternative explanations, even when using widespread classical statistical tools such as linear or logistic regression. For example, in neuroimaging, it is common to use Bayesian regression methods to study the relationship between functional magnetic resonance imaging signals against multiple independent variables that are partially correlated with each other[60–62]. Since Bayesian model evidence naturally balances goodness-of-fit against model complexity[63–65], model selection techniques can be used in such problems to make proper inferences about whether adding or removing a potential independent variable improves the model or not. The proposed method for assessing power is thus broadly applicable to any statistical problem where Bayesian model selection is useful, not just computational modelling studies. The strength of Bayesian model selection lies in its ability to evaluate and compare multiple competing explanations or models simultaneously, transcending the limitations of classical null hypothesis testing.

Research practices in diverse fields from criminology, to social psychology, biomedical research and neuroscience have faced severe criticism due to the low reproducibility of their findings[66–76]. The underlying causes of these research problems could potentially create similar issues in computational studies of human behaviour. One major criticism of these fields, especially in neuroscience, over the past decade has been the lack of sufficient statistical power given the likely effect sizes[33,77–80], a pitfall that computational behavioural studies must avoid. Low-powered studies, by definition, have a reduced chance of detecting true effects[32]. Moreover, and perhaps more importantly, when they do detect effects, they tend to exaggerate their magnitude. This effect inflation is particularly likely when discoveries are based on applying selection thresholds to statistics, whether frequentist or Bayesian[33,34], as is common in model selection studies. Moreover, the use of fixed effects methods for model selection can exacerbate this issue due to their extreme sensitivity to outliers, potentially leading to a 'winner's curse' effect[34,81] where evidence in favour of a (usually more complex) model of interest is inflated due to extreme effects in a small number of subjects, as shown in our simulation analyses.

This study highlights the importance of considering both sample size and model space when conducting model selection. The statistical framework we have developed provides researchers with a valuable tool for power analysis and underscores the need for greater attention to statistical power in the field. By addressing these challenges, we can strengthen the foundations of computational modelling research and advance our understanding of human behaviour and cognition.

Methods

Algorithm

We now detail the algorithm for calculating statistical power in model selection studies. Let K denote the number of models and N the sample size. We set the Dirichlet prior parameter $\mathbf{c}$, a 1-by-K vector with all

elements equal to c, and the acceptable false positive rate α. We also set the Dirichlet parameter for the true distribution to $\mathbf{b}$, a 1-by-K vector with all elements equal to b. L is the number of simulations. The algorithm contains two stages.

Stage 1: calculating the decision threshold. Generate L simulations under the null scenario, where no model is more likely at the population level. For each simulation l:

- Draw group sample $\mathbf{g}^l \sim \text{Multinomial}(N, \mathbf{p})$, where $\mathbf{p} = [1/K, \ldots, 1/K]$ is a 1-by-K vector and '~' denotes 'distributed as'.
- Calculate exceedance probabilities for all models, $\varphi_k^l = \Pr\left[m_k > m_j, \forall j \neq k | \mathbf{g}^l + \mathbf{c}\right]$, where $k = 1, \ldots, K$ and $\mathbf{g}^l + \mathbf{c}$ is the vector of Dirichlet parameters.
- Record $\varphi_{\max}^l = \max_k \varphi_k^l$.

Determine φ_{crit} as the minimum threshold for which $\frac{1}{L}\sum_{l=1}^{L} \mathbb{1}\left[\varphi_{\max}^l \geq \varphi_{\text{crit}}\right] \leq \alpha$, where $\mathbb{1}[\text{condition}]$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

Stage 2: power calculation. Generate L simulations where model 1 is the true model at the population level. For each simulation l:

- Keep drawing $\mathbf{m}^l \sim \text{Dir}(\mathbf{b})$, where $\mathbf{m}^l = [m_1^l, \ldots, m_K^l]$ until $m_1^l > m_k^l$ for all $k \neq 1$.
- Generate group sample $\mathbf{g}^l \sim \text{Multinomial}(N, \mathbf{m}^l)$, where $\mathbf{g} = [g_1, \ldots, g_K]$ is a 1-by-K vector.
- Calculate exceedance probability for model 1, $\varphi_1^l = \Pr\left[m_1 > m_k, \forall k \neq 1 | \mathbf{g}^l + \mathbf{c}\right]$, where $\mathbf{g}^l + \mathbf{c}$ is a 1-by-K vector of Dirichlet parameters.

Calculate power as $\frac{1}{L}\sum_{l=1}^{L} \mathbb{1}\left[\varphi_1^l > \varphi_{\text{crti}}\right]$.

Simulation analyses: for the results presented in Figs. 1 and 2, we used this algorithm with $b = c = 1$, $L = 10{,}000$ and $\alpha = 0.05$. Exceedance probabilities for a given set of Dirichlet parameters were computed by drawing 10,000 samples from the corresponding Dirichlet distribution.

Narrative review

To characterize common Bayesian model selection practices in computational modelling studies, we conducted a selective narrative literature review[82]. We focused on behavioural and neural studies involving human participants published in three top-tier journals: *Nature Human Behaviour*, *Nature* and *Science*. We searched these journals' websites using various combinations of terms related to Bayesian model selection. Our initial search criteria included: (1) publications from 2018 to July 2024, (2) original research papers only and (3) subject fields limited to psychology and neuroscience. We then examined the resulting studies to include only those that: (1) involved human participants and (2) used Bayesian tools specifically for model selection, excluding those papers that merely used Bayesian regression for data analysis without explicit model selection between competing models. This process yielded 54 studies for inclusion in our review. Two studies with sample sizes of only five were excluded from the analyses, as their limited recruitment probably reflected practical constraints rather than standard study design. The complete list of these studies is available on the GitHub repository for this study. The author then evaluated how these studies performed model selection, and what was their sample size and model space size. For studies reporting Bayesian model selection across multiple experiments or different groups of participants, we calculated power separately for each reported analysis and used the average of these values in Fig. 2. However, if model selection was conducted on participants from all experiments combined, we used the total number of participants as the sample size.

Simulation analyses of the fixed effects approach

Simulation analyses of the fixed effects approach were conducted using 10,000 simulations, where in each simulation 200 data points for 50 subjects were generated across three models. In the simulation presented in Table 1, we designed a scenario with no difference between models, randomly drawing the probability of all data points for all 50 subjects from a uniform distribution for each model. For simulation 1 presented in Table 2, we generated 49 synthetic subjects similar to the previous simulation, plus one outlier subject. Simulation 2 presented in Table 2 involved a two-step process: first, we randomly generated data points between 0 and 1, then discarded simulations where any model other than model 2 was the fixed effects winner (using a threshold of 3), repeating this process until we accumulated 10,000 simulations with model 2 as the clear winner. For outlier analyses presented in Tables 1 and 2, we added an extreme outlier subject to each simulation, randomly generating the probability of all data points for model 1 between 0.9 and 1, while randomly drawing probabilities for the other models between 0 and 1. The simulation presented in Fig. 3 was similar to the one presented in Table 2, with the only difference being that the outlier was designed to be modest, thus its per-trial probability for model 1 was only slightly biased, that is, between 0.1 and 1. In all simulation analyses, we quantified how often each model was declared as winner by the fixed effects model selection process, using a threshold of 3 for the difference between log-evidence (that is, equivalent to exp(3) as the threshold for Bayes factor).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

No data were collected for this study. Source data are provided with this paper.

Code availability

All analyses were conducted using custom code written in MATLAB v9.13.0 (R2022b). The analysis code is available via GitHub at https://github.com/payampiray/power_computational_studies. Python implementation of the power analysis method introduced here is available via GitHub at https://github.com/payampiray/cbm_power.

References

1. Allen, K. et al. Using games to understand the mind. *Nat. Hum. Behav.* **8**, 1035–1043 (2024).
2. Buhrmester, M. D., Talaifar, S. & Gosling, S. D. An evaluation of Amazon's mechanical turk, its rapid rise, and its effective use. *Perspect. Psychol. Sci.* **13**, 149–154 (2018).
3. Doerig, A. et al. The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
4. Egnor, S. E. R. & Branson, K. Computational analysis of behavior. *Annu. Rev. Neurosci.* **39**, 217–236 (2016).
5. Gillan, C. M. & Daw, N. D. Taking psychiatry research online. *Neuron* **91**, 19–23 (2016).
6. Guest, O. & Martin, A. E. How computational modeling can force theory building in psychological science. *Perspect. Psychol. Sci.* **16**, 789–802 (2021).
7. Huys, Q. J. M., Browning, M., Paulus, M. P. & Frank, M. J. Advances in the computational understanding of mental illness. *Neuropsychopharmacology* **46**, 3–19 (2021).
8. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
9. Stewart, N., Chandler, J. & Paolacci, G. Crowdsourcing samples in cognitive science. *Trends Cogn. Sci.* **21**, 736–748 (2017).
10. Cohen, J. D. et al. Computational approaches to fMRI analysis. *Nat. Neurosci.* **20**, 304–313 (2017).
11. Palminteri, S., Wyart, V. & Koechlin, E. The importance of falsification in computational cognitive modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).

12. Gillan, C. M. et al. Comparison of the association between goal-directed planning and self-reported compulsivity vs obsessive-compulsive disorder diagnosis. *JAMA Psychiatry* **77**, 77–85 (2020).
13. Hunter, L. E., Meer, E. A., Gillan, C. M., Hsu, M. & Daw, N. D. Increased and biased deliberation in social anxiety. *Nat. Hum. Behav.* **6**, 146–154 (2022).
14. van Timmeren, T., Piray, P., Goudriaan, A. E. & van Holst, R. J. Goal-directed and habitual decision making under stress in gambling disorder: an fMRI study. *Addict. Behav.* **140**, 107628 (2023).
15. Piray, P. & Daw, N. D. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nat. Commun.* **12**, 4942 (2021).
16. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**, 1148–1160 (2018).
17. Mattar, M. G. & Lengyel, M. Planning in the brain. *Neuron* **110**, 914–934 (2022).
18. Piray, P. & Daw, N. D. Reconciling flexibility and efficiency: medial entorhinal cortex represents a compositional cognitive map. *Nat. Commun.* **16**, 7444 (2025).
19. Zorowitz, S., Solis, J., Niv, Y. & Bennett, D. Inattentive responding can induce spurious associations between task behaviour and symptom measures. *Nat. Hum. Behav.* **7**, 1667–1681 (2023).
20. Pitt, M. A. & Myung, I. J. When a good fit can be bad. *Trends Cogn. Sci.* **6**, 421–425 (2002).
21. Steyvers, M., Lee, M. D. & Wagenmakers, E.-J. A Bayesian analysis of human decision-making on bandit problems. *J. Math. Psychol.* **53**, 168–179 (2009).
22. Berger, J. O. & Pericchi, L. R. The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.* **91**, 109–122 (1996).
23. Dienes, Z. Using Bayes to get the most out of non-significant results. *Front. Psychol.* **5**, 781 (2014).
24. Kruschke, J. K. Bayesian estimation supersedes the *t* test. *J. Exp. Psychol. Gen.* **142**, 573–603 (2013).
25. Lee, M. D. & Wagenmakers, E.-J. Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychol. Rev.* **112**, 662–668 (2005).
26. Wagenmakers, E.-J. A practical solution to the pervasive problems of *p* values. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
27. Daunizeau, J., Adam, V. & Rigoux, L. VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput. Biol.* **10**, e1003441 (2014).
28. Daw, N. D. in *Decision Making, Affect, and Learning: Attention and Performance XXIII* (eds Delgado, M. R. et al.) 3–38 (Oxford Univ. Press, 2011).
29. Kass, R. E. & Raftery, A. E. Bayes factor. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
30. Keysers, C., Gazzola, V. & Wagenmakers, E.-J. Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nat. Neurosci.* **23**, 788–799 (2020).
31. Wilson, R. C. & Collins, A. G. Ten simple rules for the computational modeling of behavioral data. *eLife* **8**, e49547 (2019).
32. Moher, D., Dulberg, C. S. & Wells, G. A. Statistical power, sample size, and their reporting in randomized controlled trials. *JAMA* **272**, 122–124 (1994).
33. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
34. Ioannidis, J. P. A. Why most discovered true associations are inflated. *Epidemiology* **19**, 640–648 (2008).
35. Stephan, K. E., Weiskopf, N., Drysdale, P. M., Robinson, P. A. & Friston, K. J. Comparing hemodynamic models with DCM. *NeuroImage* **38**, 387–401 (2007).
36. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *NeuroImage* **46**, 1004–1017 (2009).
37. Stephan, K. E. et al. Ten simple rules for dynamic causal modeling. *NeuroImage* **49**, 3099–3109 (2010).
38. Penny, W. D. et al. Comparing families of dynamic causal models. *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1000709 (2010).
39. Piray, P., Dezfouli, A., Heskes, T., Frank, M. J. & Daw, N. D. Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Comput. Biol.* **15**, e1007043 (2019).
40. Rigoux, L., Stephan, K. E., Friston, K. J. & Daunizeau, J. Bayesian model selection for group studies—revisited. *NeuroImage* **84**, 971–985 (2014).
41. Penny, W. D. and Holmes, A. J. in *Statistical Parametric Mapping: The Analysis of Functional Brain Images* 1st edn (eds Friston, K. J. et al.) 156–165 (Elsevier, 2007).
42. Klaassen, F., Zedelius, C. M., Veling, H., Aarts, H. & Hoijtink, H. All for one or some for all? Evaluating informative hypotheses using multiple *N* = 1 studies. *Behav. Res. Methods* **50**, 2276–2291 (2018).
43. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
44. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
45. Watanabe, S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010).
46. Friston, K. J., Mattout, J., Trujillo-Barreto, N., Ashburner, J. & Penny, W. Variational free energy and the Laplace approximation. *NeuroImage* **34**, 220–234 (2007).
47. Beckmann, C. F., Jenkinson, M. & Smith, S. M. General multilevel linear modeling for group analysis in FMRI. *NeuroImage* **20**, 1052–1063 (2003).
48. Friston, K. J., Holmes, A. P., Price, C. J., Büchel, C. & Worsley, K. J. Multisubject fMRI studies and conjunction analyses. *NeuroImage* **10**, 385–396 (1999).
49. Holmes, A. P. & Friston, K. J. Generalisability, random effects and population inference. *NeuroImage* **7**, S754 (1998).
50. Wagenmakers, E.-J., Ratcliff, R., Gomez, P. & Iverson, G. J. Assessing model mimicry using the parametric bootstrap. *J. Math. Psychol.* **48**, 28–50 (2004).
51. Cohen, J. Statistical Power Analysis for the Behavioral Sciences (Routledge, 2009). .
52. Huys, Q. J. M. et al. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput. Biol.* **8**, e1002410 (2012).
53. Wiecki, T. V., Sofer, I. and Frank, M. J. HDDM: hierarchical Bayesian estimation of the drift-diffusion model in Python. *Front. Neuroinform.* https://doi.org/10.3389/fninf.2013.00014 (2013).
54. Adams, R. A., Huys, Q. J. M. & Roiser, J. P. Computational psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* **87**, 53–63 (2016).
55. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
56. Piray, P. et al. Impulse control disorders in Parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. *J. Neurosci.* **34**, 7814–7824 (2014).
57. Piray, P., Ly, V., Roelofs, K., Cools, R. & Toni, I. Emotionally aversive cues suppress neural systems underlying optimal learning in socially anxious individuals. *J. Neurosci.* https://doi.org/10.1523/JNEUROSCI.1394-18.2018 (2019).
58. Nassar, M. R. & Frank, M. J. Taming the beast: extracting generalizable knowledge from computational models of cognition. *Curr. Opin. Behav. Sci.* **11**, 49–54 (2016).

59. Piray, P. & Daw, N. D. Computational processes of simultaneous learning of stochasticity and volatility in humans. *Nat. Commun.* **15**, 9073 (2024).
60. Friston, K. et al. Bayesian decoding of brain images. *NeuroImage* **39**, 181–205 (2008).
61. Woolrich, M. W. et al. Bayesian analysis of neuroimaging data in FSL. *NeuroImage* **45**, S173–S186 (2009).
62. Woolrich, M. W. Bayesian inference in FMRI. *NeuroImage* **62**, 801–810 (2012).
63. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
64. MacKay, D. J. C. Bayesian interpolation. *Neural Comput.* **4**, 415–447 (1992).
65. MacKay, D. J. C. Information Theory, Inference, and Learning Algorithms (Cambridge Univ. Press, 2003).
66. Al-Shahi Salman, R. et al. Increasing value and reducing waste in biomedical research regulation and management. *Lancet* **383**, 176–185 (2014).
67. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).
68. Chalmers, I. et al. How to increase value and reduce waste when research priorities are set. *Lancet* **383**, 156–165 (2014).
69. Fanelli, D. 'Positive' results increase down the Hierarchy of the Sciences. PLoS ONE 5, e10068 (2010).
70. Glasziou, P. et al. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **383**, 267–276 (2014).
71. Ioannidis, J. P. A., Tarone, R. & McLaughlin, J. K. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology* **22**, 450–456 (2011).
72. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532 (2012).
73. Macleod, M. R. et al. Biomedical research: increasing value, reducing waste. *Lancet* **383**, 101–104 (2014).
74. Makel, M. C., Plucker, J. A. & Hegarty, B. Replications in psychology research: how often do they really occur? *Perspect. Psychol. Sci.* **7**, 537–542 (2012).
75. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
76. Pridemore, W. A., Makel, M. C. & Plucker, J. A. Replication in criminology and the social sciences. *Annu. Rev. Criminol.* **1**, 19–38 (2018).
77. Carp, J. The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage* **63**, 289–300 (2012).
78. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
79. Poldrack, R. A. et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. *Nat. Rev. Neurosci.* **18**, 115–126 (2017).
80. Yarkoni, T. Big correlations in little studies: inflated fMRI correlations reflect low statistical power-commentary on Vul et al. (2009). *Perspect. Psychol. Sci.* **4**, 294–298 (2009).
81. Lindstromberg, S. The winner's curse and related perils of low statistical power—spelled out and illustrated. *Res. Methods Appl. Linguist.* **2**, 100059 (2023).
82. Grant, M. J. & Booth, A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info. Libr. J.* **26**, 91–108 (2009).

Acknowledgements

This work was supported by grant no. R21MH134217 from the National Institute of Mental Health. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

P.P. was responsible for conceiving the research idea, developing the theoretical framework, performing the analyses and writing the manuscript.

Competing interests

The author declares no competing interests.

Additional information

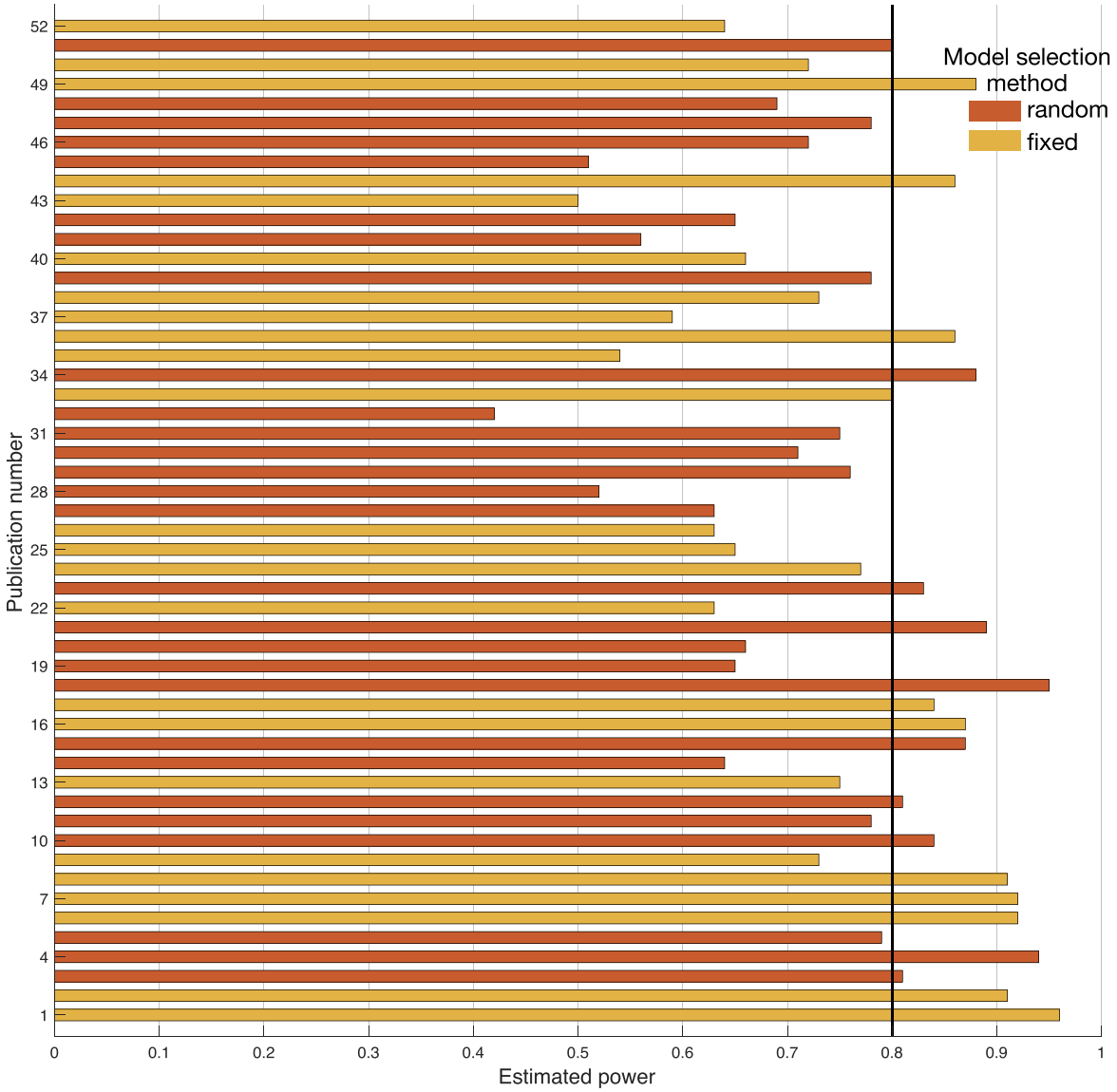
Extended data is available for this paper at https://doi.org/10.1038/s41562-025-02348-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02348-6.

Correspondence and requests for materials should be addressed to Payam Piray.

Peer review information *Nature Human Behaviour* thanks Frederik Aust, Min Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-025-02348-6



Extended Data Fig. 1 | Narrative review of the literature with fixed medium effect size across studies. Estimated power for the 52 reviewed studies is plotted, with the target effect size fixed at 0.3. The y-axis indicates publication numbers in the same order as in Fig. 2. This effect size corresponds to Cohen's g = 0.15 for standard binomial tests (assessing deviations from a 0.5 probability across two categories), conventionally considered a medium effect[51]. The algorithm used here was the same as in the main text, except that for this analysis the Dirichlet parameter *b* was calibrated so that the true effect size (the mean difference between the best and the second-best models across samples) was within ± 0.01 of the target value for each study. This analysis indicates that 31 of the 52 studies fell below the standard 0.8 power threshold. Unlike the main analysis, however, fixing effect sizes substantially increases the overall power, although the majority of studies still do not reach the 0.8 threshold. Given that several studies fall near the 0.8 threshold (the black vertical line), this analysis was repeated with ten different random seeds for those studies, and the mean power is reported. The standard errors of the mean are zero at the resolution of the computed power (two decimal places). The reported values are conservative, as all computed power values were rounded upward to two decimal places.

nature portfolio

Corresponding author(s): Payam Piray
Last updated by author(s): Sep 30, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data were collected for the current study.
Data analysis	All analyses were conducted using custom code written in MATLAB v9.13.0 (R2022b). The analysis code is available at https://github.com/payampiray/power_computational_studies. Python implementation of the power analysis method introduced here is available at https://github.com/payampiray/cbm_power

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

No data were collected for the current study.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods used to predetermine sample size, because all results are based on computer simulations, and no human or animal subject or sample was recruited. For the narrative review analysis, we focused on behavioral and neural studies involving human participants published in three top-tier journals: Nature Human Behaviour, Nature, and Science. The initial criteria restricted the search to original research articles in psychology and neuroscience. From these, we included only studies that involved human participants and explicitly applied Bayesian methods for model selection, excluding those that used Bayesian regression without model comparison. This process identified 54 studies, with two excluded due to extremely small sample sizes, likely determined by factors other than power analysis.
Data exclusions	For the narrative review analysis, two studies were excluded because they had very small sample sizes (n = 5), which were likely determined by factors other than power analysis. No data were excluded from the simulation analyses.
Replication	Given that all results are based on computer simulations, reproducibility is guaranteed. Each simulation experiment was repeated thousand of times and summary statistics (with negligible errorbars) are reported. Moreover, as this study is based entirely on computer simulations, covariates are not applicable. Accordingly, the method provides a general solution to the power analysis problem that does not depend on study-specific covariates.
Randomization	Given that all results are based on computer simulations, no randomization was needed.
Blinding	Given that all results are based on computer simulations, no blinding was needed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.