What the Average Really Means: Dissociating Effect Size and Effect Prevalence using *p*-curve Mixtures

John P. Veillette[1†]
Howard C. Nusbaum[1]

[1]Department of Psychology, University of Chicago
[†]Correspondence should be addressed to John P. Veillette; E-mail: johnv@uchicago.edu

Author Contributions

H.C.N.: Funding acquisition, Resources, Supervision, and Writing - review & editing. **J.P.V.:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.

Acknowledgements

This project was supported by NSF BCS 2024923 to H.C.N.; J.P.V. was supported by NSF DGE 1746045. This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

Abstract

Most research in the behavioral sciences aims to characterize effects of interest using sample means intended to describe the "typical" person. A difference in means is usually construed as a size difference in an effect common across subjects. However, mean effect size varies with both *within-subject effect size* and *population prevalence* (proportion of population showing the effect) in compared groups or across conditions. Few studies consider how prevalence affects mean effect size measurements and existing estimators of prevalence are, conversely, confounded by uncertainty about within-subject power. We introduce a widely applicable Bayesian method, the *p*-curve mixture model, that jointly estimates prevalence and effect size. Our approach outperforms existing prevalence estimation methods when within-subject power is uncertain and is sensitive to differences in prevalence or effect size across groups or experimental conditions. We present examples, extracting novel insights from existing datasets, and provide a user-facing software tool.

Introduction

Many psychology and cognitive neuroscience studies support their claims by showing that there is a statistically significant difference in the mean size of some effect – say, a difference in some variable as the result of an intervention – between two or more conditions or groups. Such a difference is often interpreted as a change in the effect size's magnitude between groups. This interpretation follows from a set of typical, though usually unstated, assumptions: (1) a group condition mean characterizes the condition response of a "group-representative subject," (2) measurements from actual subjects represent noise around this central tendency, and (3) subjects who are very far from the mean are "outliers" who may safely (or even should) be discarded from analyses because they have systematic characteristic differences from the representative hypothetical subject being investigated. However, these assumptions are not always correct and, we argue, can therefore obscure meaningful patterns in sample variance that may be present in subpopulations, even if not in a modal subject.

Population prevalence refers to the proportion of a population (or sample from that population) demonstrating an effect. In other words, the mean effect size in a population is, notably, only equal to the *within-subject* effect size if the *population prevalence* of the effect is 100%. If an effect is heterogenous – some people show it and some do not – then the sample mean will be "watered down" relative to the within-subject effect size, and will thus be representative of neither the subgroup that shows the effect nor that which does not. In an experiment, the mean effect size could vary across groups or conditions because the within-subject effect size differs, but it could also vary if the *proportion of people who show the effect* differs.

This distinction bears important implications. For example, an intervention that has a strong within-subject effect, even if it only applies to a modest subpopulation, could still have useful practical or clinical applications. Conversely, an intervention that achieves an equivalent group mean effect size when averaging over a negligible within-subject effect that is widely present in the population may be less useful in practice – but also more likely to replicate in a new sample, as a researcher does not have to be "lucky" in sampling subjects from the selected subgroup. While may be tempting to assume such situations are exceptions, the strong correlation between group mean effect size and the between-study heterogeneity of that effect among preregistered replications indicates the largest effects are actually the most likely to vary across subjects (Olsson-Collentine et al., 2020). Indeed, there is a growing recognition across fields of behavioral research that effects measured and mechanistic models fit at the group-level often deviate markedly from what is observed in any single subject (Bolger et al., 2019; Botella et al., 2019; Grandy et al., 2017; Moreau & Corballis, 2019; Navarro et al., 2006). As such, Bryan and colleagues have recently argued that the antidote to the reproducibility crisis in the behavioral sciences is a "heterogeneity revolution," in which systematic approaches to sampling and to quantifying population heterogeneity are adopted (Bryan et al., 2021). While we agree with Bryan et al. (2021) that researchers should more often adopt sampling plans that explicitly aim to capture population heterogeneity when measuring already-established effects, much research aims merely to establish the existence of novel effects predicted by theories. Even for the latter sort of research, it would be highly informative for empirical studies to estimate and report the proportion of the (sampled) population to which observed effects can be expected to apply, as well as useful

quantities such as within-subject effect size or power estimates, both of which can inform sampling approaches for future confirmatory or applied research.

Currently, standard approaches to dealing with population heterogeneity are likely to exacerbate this problem; for instance, outlier removal aims to eliminate subgroup differences before computing descriptive or inferential statistics. This approach yields sample mean effect size estimates that neither are good estimates of the true population mean, as they exclude parts of the population, nor capture real heterogeneity. A better approach would be to jointly estimate the population prevalence of observed effects and the within-subject effect size for those to whom the effect applies, so as to quantify heterogeneity instead of artificially removing it. Moreover, a systematic difference in the prevalence of an effect between two populations (i.e. between-group difference) may be of scientific interest; for example, as psychological differences between people from Western and non-Western cultures have become increasingly well-documented, it would be fruitful to dissociate whether such differences reflect the prevalence or size of effects, possibilities that suggest distinct theoretical explanations (Henrich et al., 2010; Muthukrishna et al., 2020). Moreover, one might wish to estimate the difference in the prevalence of two effects in the same population (i.e. within-group difference) or the conditional probability/prevalence of one effect given that a person shows a different effect – which may be used as evidence that two behavioral and/or neural phenotypes are driven by the same latent trait.

When the ground truth status of individuals is known, estimating prevalence is simply a matter of computing the sample proportion of individuals showing the effect. In practice, however, we must deal with the possibility that some subjects may show an effect in reality, but we failed to detect it in our experiment. In such cases, sample proportions are systematically lower than the population prevalence, so frequentist approaches to prevalence inference have primarily aimed to put a lower bound on population prevalence, rather than an estimate (Allefeld et al., 2016; Donhauser et al., 2018). Without such an estimate, however, one cannot easily compare groups or conditions. Ince and colleagues proposed a simple but powerful Bayesian method in which, given the binary outcomes of null hypothesis significance tests (NHST) conducted within each subject, models the incidence of a significant within-subject test as a Binomial distribution as a function of the population prevalence and within-subject power (Ince et al., 2021). Not only does this approach allow prevalence to be compared across groups or conditions, but it is applicable whenever a *p*-value can be computed for each subject. However, estimates are sensitive to the arbitrary significance level of the within-subject test, and as we show here, the approach cannot be fruitfully extended to settings in which the within-subject power/effect size is unknown or may vary across groups or conditions. That is, the Binomial model cannot simultaneously estimate both population prevalence of an effect and the within-subject effect size.

In our past work, we have dealt with this issue by using Bayesian mixture modeling (Hedger et al., 2020; Veillette, Gao, et al., 2024). This approach models the distribution of the subject-level observations as a mixture of two or more subgroup distributions, allowing one to infer the parameters of these distributions and the proportion of subjects that belong to each subgroup simultaneously. However, using this approach for prevalence estimation requires constructing an appropriate Bayesian model of the data, which may require substantial expertise or might even be intractable. In the present work, we aim to combine the advantages of mixture modeling with the broad applicability of the Binomial model. Our "*p*-curve mixture model"

simultaneously infers the population prevalence and the (relative) within-subject effect size from the unthresholded *p*-values of within-subject null hypothesis significance tests. Thus, it can be uniformly applied to any study in which a significance test can be performed per-subject. Since it does not require assuming a fixed power *a priori* as do previous methods, our method can be applied to datasets that were not originally collected with within-subject statistics in mind; this wide applicability affords the opportunity for researchers to estimate population heterogeneity using existing data and experimental designs.

Methods

1. Bayesian mixture models for prevalence estimation

We will first describe how Bayesian mixture models can be used to estimate prevalence and effect size generally, before moving on to *p*-curve mixture models which are a special case. An "effect" here may, for example, refer to that of a treatment of intervention, or to a quantifiable behavioral characteristic/skill – such as a performance measure on a task – that can be compared to a meaningful null ("no effect") distribution, as in the example in this section.

The general Bayesian approach to statistics is to first describe a generative model of one's data – that is, to specify the likelihood of the data given some parameters – and then use Bayes' rule to infer plausible values for the parameters given the observed data. In a Bayesian mixture model, this generative model entails that an observation can come from any one of at least two distributions; thus, the marginal likelihood of the data is a mixture of the likelihoods of the component distributions, weighted by the probabilities of an observation coming from each component. In most applications of Bayesian Mixture models, the parameters of *all* component distributions (e.g. their mean and scale) and the relative probabilities of each component distribution are all estimated during model fitting. This is what most mixture modeling software packages, Bayesian or frequentist, implement out-of-the-box (Benaglia et al., 2010). Mixture models, Bayesian or otherwise, have often been used as a sort of probabilistic clustering technique to infer the presence of latent subgroups within a large-sample datasets (Kim et al., 2020; Lubke & Muthén, 2005), or to infer effect heterogeneity across studies in meta-analyses (Moreau & Corballis, 2019). However, mixture models can be applied to achieve more specific inferential goals by placing constraints on the shape of the component distributions (Frischkorn & Popov, 2023).

When one component distribution is constrained such that one component distribution is a null H_0 distribution – the likelihood of a subjects' data if there is no effect – and the other distribution is an alternative H_1 distribution which ideally incorporates prior information about the range of plausible effect sizes under the alternative hypothesis, then the probability assigned to the H_1 distribution is an estimate of H_1's prevalence in the sampled population.

One such generative model for accuracy data, which we have used in previous work and in Example A, is written as follows (Veillette, Gao, et al., 2024). Subject i, who is participating in a two-alternative forced choice task, answers correctly with accuracy κ_i, such that their total number of correct trials over the task is $k_i \sim \text{Binomial}(n_{\text{trials}}, \kappa_i)$. Subjects perform above chance

at this task (i.e. H_1 is true) with probability γ, the population prevalence. If H_1 is indeed true for subject i, then $2(\kappa_i - 0.5) \sim \text{Beta}(\alpha, \beta)$, which constrains κ_i to be between 0.5 and 1 (the definition of an above-chance accuracy), and $\kappa_i = 0.5$ (i.e. chance accuracy) if H_0 is true. For Bayesian estimation, priors must be placed on the parameters of the beta distribution α, β and on the prevalence γ, but these priors can be weakly or non-informative, such as a $\gamma \sim \text{Uniform}(0, 1)$ prior for H_1 prevalence.

Once the model and prior distributions are specified, the posterior distribution for each parameter – representing beliefs about plausible parameter values after conditioning on the observed data – is given by Bayes' rule. A normalizing constant for the posterior distribution (i.e. to make it sum to 1) for a given parameter cannot usually be derived in closed form, but the posterior can be sampled from without knowing the normalizing constant using Markov-Chain Monte Carlo (Hoffman & Gelman, 2014).

To specify the above model, we had to (a) know what the likelihood of the data was under the null hypothesis (e.g. $k_i \sim \text{Binomial}(n_{\text{trials}}, 0.5)$), (b) know what the likelihood of the data was, given some effect size (e.g. accuracy) was under the alternative hypothesis, (c) have a prior describing plausible values for that effect size (e.g. between 0.5 and 1), and (d) know how to program the custom model and approximate posterior distributions for its parameters in a probabilistic programming framework such as *PyMC* or *Stan* that will perform posterior sampling for us (Gelman et al., 2015; Patil et al., 2010). We simply do not always have these ingredients. Often, even the shape of the null distribution is not known a priori – e.g. the same situations where one might use a non-parametric test in the NHST framework. Moreover, it is not practical to assume all researchers have the combination of domain and technical knowledge to implement such an ad hoc statistical model for their own, idiosyncratic dataset. If there were, however, a way to transform observed data from arbitrary distributions into a distribution we always know enough about to model this way – much like users of NHST can almost always resort to a non-parametric test – then researchers would not need to write their own software for each use case.

2. The *p*-curve mixture model

2.1. p-curves

Since *p*-values most commonly appear in the setting of null hypothesis significance testing (NHST), it may be unconventional to think of them as random variables. Indeed, in studies in which only one *p*-value is computed or multiple *p*-values are computed on the same data, thus are not independent observations, it would not be useful to think of them this way. However, when a test is repeated multiple times on *independent* samples, then the independent *p*-values are indeed random variables with a probability distribution. This sort of distribution, called a *p*-curve, has been leveraged to study and correct for the effects of publication bias in the meta-analysis literature (Simonsohn et al., 2014a, 2014b, 2015).

Under the null hypothesis, a *p*-curve is always Uniform(0, 1); this fact is built into the logic of NHST, since the *p*-value can only fall under the significance level α just α proportion of the time for *any* arbitrary α. Under the alternative hypothesis, the *p*-curve becomes increasingly left skewed the larger the effect size or sample size (see Figure 1). One can derive an exact *p*-curve for

a given statistical test, such as the *t*-test, as a function of the effect size and sample size (Simonsohn et al., 2014a). Usefully, such an approach can be used to recover an effect size merely from repeated *p*-values as has been done in meta-analysis (Simonsohn et al., 2014b).

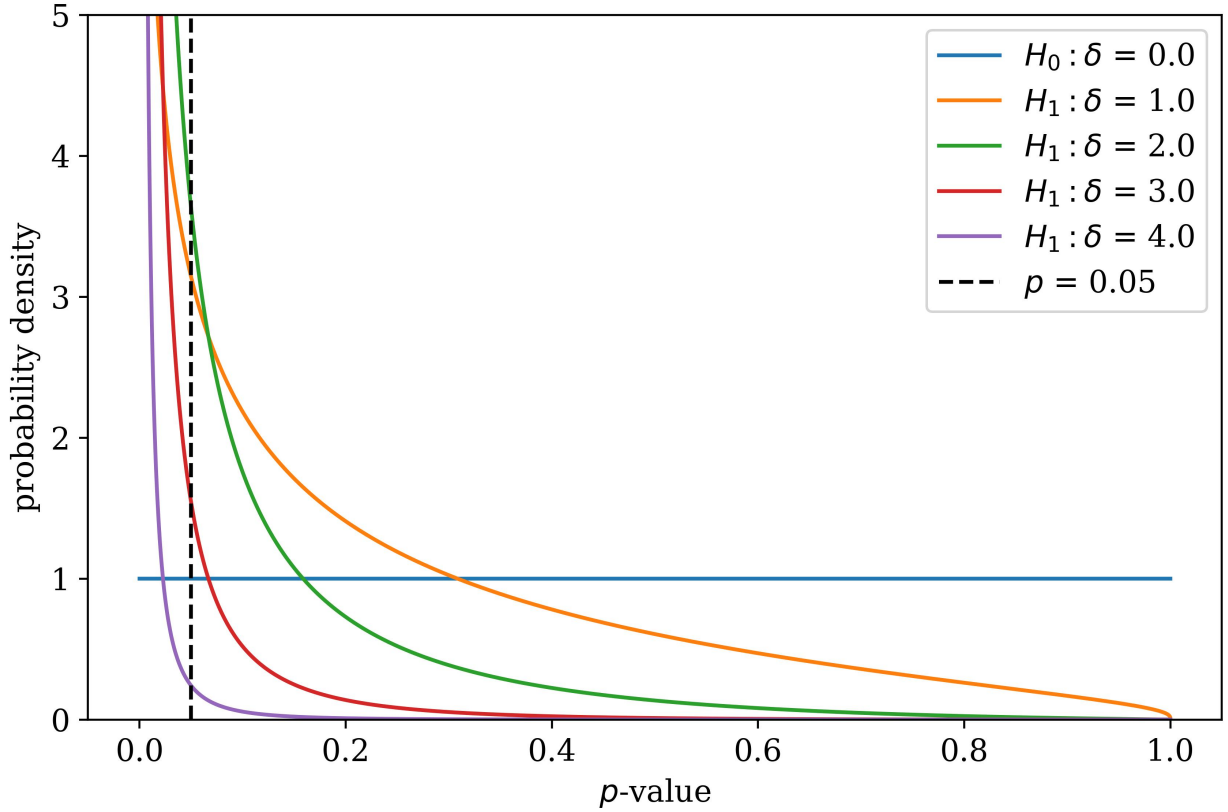

Figure 1: A probability distribution for *p*-values. Describing the likelihood of *p*-values from repeated, independent tests of an effect with size δ, these *p*-curves – specified by Equation 1 – are valid probability distributions that integrate to one over the interval (0, 1). When the null hypothesis is true ($\delta = 0$), the *p*-curve is equivalent to the standard uniform distribution.

However, since the "sample size" is not always well-defined for a within-subject study (e.g. consecutive trials in a task or neuroimaging measurement may not be independent observations), we specify a general *p*-curve in Equation 1 that depends on a single effect size parameter δ using the standard normal probability density function φ and cumulative distribution function ϕ.

$$p \sim f(p, \delta) \text{ where } f(p, \delta) = \frac{\varphi(\phi^{-1}(1-p) - \delta)}{\varphi\left(\phi^{-1}(1-p)\right)} \quad (1)$$

The use of the normal distribution in our specification does not entail that our model will work only with normally distributed data – unless you happen to be using a Z-test with sample size one, in which case this *p*-curve is indeed exact – as we will verify in our simulations. It should be duly noted though that, without a sample size parameter, the "effect size" parameter in our

model is abstract/unitless and should only be compared across identical experiments (same number of trials, design, etc.). However, this unitless effect size parameter can be converted back into meaningful units by taking the calculating the area under the *p*-curve distribution up to (i.e. cumulative distribution function at) some $p = \alpha$, which is interpreted as the power of a within-subject NHST at significance level α.

2.2. Mixtures of p-curves

Now that we have a probability distribution that can describe how *p*-values are distributed given an effect size, we can incorporate it into a mixture model – which can then be applied whenever per-subject *p*-values can be computed using NHST within each subject, regardless of the original distribution of the data. As in our previous mixture model, subjects show an effect (i.e. H_1 is true) with probability/prevalence γ. If H_1 is true for subject i, then their *p*-value is distributed according to $p_i \sim f(p_i, \delta)$, as defined in Equation 1, for some unitless effect size δ. If instead H_0 is true, then $p_i \sim f(p_i, 0)$ or equivalently $p_i \sim \text{Uniform}(0,1)$.

Before performing Bayesian inference given a set of observed *p*-values, one needs to assign prior distributions to γ and δ. In all simulations and examples in this manuscript, we will use an uninformative $\gamma \sim \text{Uniform}(0,1)$ prior for the prevalence of H_1 and a weakly informative prior of $\delta \sim \text{Exponential}(2/3)$. The latter was chosen as it results in a prior 90% highest-density interval for within-subject power (at significance level $\alpha = 0.05$) of roughly [0.05, 0.95], and is this relatively close to a uniform prior over within-subject power. Note that a flat prior, instead of an exponential prior, would allow unreasonably high values of δ and would not be "uninformative" in this case. Our exponential prior, per our simulations, performs well across a variety of cases.

2.3. Between-group differences

p-curve mixtures can be used to estimate the difference in prevalence and in (unitless) within-subject effect size between two independent groups of subjects. This is useful, for example, in a between-subject experimental design with random assignment, or when comparing experimental results between two distinct populations (e.g. men and women, liberal and conservatives, Westerner and Easterners, etc.).

If the groups are independent, and thus group differences can be treated as fixed effects, then it is appropriate to estimate the difference by fitting *p*-curve mixture models to each of the two populations separately, drawing samples from the parameters of both model posteriors (see ***Estimation and Software***), and subtracting the samples between the groups' models to approximate samples from the posterior of the difference. We use this approach in our EEG simulations, as well as in Example B and its accompanying code notebook.

2.4. Within-group differences

Differences in prevalence and effect size can also be estimated *within* a group, if one has conducted two tests per subject with alternative hypotheses H_1 and H_2. One may find it useful to do so if they have measured the same effect in two within-subject experimental conditions (i.e. an experiment designed to test for an interaction effect) or if they want to assess whether showing one effect

makes it more or less likely that the same subject shows some other effect (e.g. whether above-chance performance on one task is more or less prevalent than above-chance performance on another task). It may even be used when a researcher wants to know if subjects who pass some manipulation check are more or less likely to show an effect of interest – an alternative to outlier rejection that does not require binary thresholding but is instead uncertainty-weighted.

In this case, unlike in the between-group case, models cannot be estimated separately for each test, since subjects are observed in both tests and thus observations are not independent. In this joint model, each subject either shows no effect in either test (denoted $k_{00} = 1$), just the H_1 effect (denoted $k_{10} = 1$), just the H_2 effect ($k_{01} = 1$), or both effects ($k_{11} = 1$). k_{00}, k_{10}, k_{01}, and k_{11} are Categorical/Multinomial(1) distributed with probabilities $\theta_{00}, \theta_{10}, \theta_{01}, \theta_{11}$, respectively. The prevalence of H_1, then, is $\gamma_1 = \theta_{10} + \theta_{11}$ and of H_2 is $\gamma_2 = \theta_{01} + \theta_{11}$. The conditional probability that H_2 is true for a subject given that H_1 is true is $P(H_2|H_1) = \theta_{11}/\gamma_1$ and vice versa. We place a minimally informative Dirichlet(1, 1, 1, 1) prior (a common default) on the θ's.

This extended model is actually somewhat difficult to implement, as the discrete latent variables k must be marginalized out analytically to ensure robust sampling from the posterior during estimation. We have provided, as for the basic *p*-curve mixture model, a user-friendly wrapper around our optimized implementation (see ***Estimation and Software***).

2.5. Estimation and Software

Posterior distributions for all parameters in a *p*-curve mixture model can be approximated by drawing thousands of samples from the posterior (we use 5,000 in our simulations and examples) using a No-U-Turn sampler or other Markov Chain Monte Carlo sampler (Hoffman & Gelman, 2014). This procedure would be laborious to program by hand, but fortunately a number of "probabilistic programming" frameworks can perform this sampling for you, which makes the *p*-curve mixture model quite simple to implement for those already familiar with Bayesian modeling (Gelman et al., 2015; Patil et al., 2010); our implementation uses *PyMC* for posterior sampling. Posterior samples can be summarized using the posterior mean/expectation (estimated as the mean of the posterior samples and the 95% (or whatever percent) highest density interval (HDI), which is the smallest interval such that there is a 95% posterior probability the estimated parameter falls within that interval.

We expect most researchers who could benefit from *p*-curve mixture modeling may not be familiar with probabilistic programming or Bayesian statistics in general. To this end, we built a user-friendly interface to our model implementation so that *p*-curve mixture models can be fit and summarized in just a few lines of code. This interface is available as a Python package called *p2prev*, which can be downloaded from the PyPI package server, from GitHub, or from Zenodo (see ***Data and Code Availability***). We provide documentation and code examples (corresponding to the examples in this paper) along with the package; some examples use the package interface (for beginners) and some implement the model directly in *PyMC* (for advanced users who may want to modify the model).

3. The Binomial model

To illustrate the advantages of *p*-curve mixture models, we also implement a version of Ince et al.'s (2021) Binomial model for estimating population prevalence (Ince et al., 2021), extended to accommodate uncertainty about within-subject power/effect size. In this model, the probability of rejecting the null hypothesis for each subject is $\pi = (1-\gamma)\alpha + \gamma(1-\beta)$, given the prevalence γ, the Type I error rate α, and the Type II error rate β of the within-subject NHST. The observed number of rejections k out of n subjects, then, is $k \sim \text{Binomial}(n, \pi)$.

Ince et al. (2021) note that, if values of α, β are fixed, then π is a deterministic function of γ. Instead of estimating the multilevel model they estimate the posterior distribution of π given observed k, which can be calculated analytically if a $\pi \sim \text{Beta}(1,1) \cong \text{Uniform}(0,1)$ prior is used for π (due to the "conjugacy" property of the beta and binomial distributions). They then solve for the implied posterior of prevalence γ algebraically. However, the Type II error rate β – a function of the within-subject effect size – is usually not known a priori. Ince et al., then, simply set $\beta = 0$ (i.e. within-subject power equals 1), such that their prevalence estimates are, in fact, a lower bound on the true prevalence – which is only a tight lower bound when statistical power is effectively 100%.

The Binomial model can easily be extended to the setting where β is unknown by putting uniform priors on both β and γ directly, then approximating their joint posterior distribution using Markov Chain Monte Carlo (Hoffman & Gelman, 2014), as we do for our *p*-curve model. Intuitively, it seems like this would yield the same benefits as our *p*-curve mixture model – namely, simultaneous estimation of population prevalence and within-subject power/effect size. However, this is not the case; the fact that the model considers the binary outcomes of the subject-level NHST as observations instead of the unthresholded *p*-values results in a likelihood in which an observed $\frac{k}{n} = 0.5$ could be (almost) equivalently well explained by $\gamma = 1, \beta = 0.5$, by $\gamma = 0.5, \beta = 0$, resulting in joint posterior distributions of γ, β that are near-symmetric about the line $\gamma = 1 - \beta$. As a result, the posterior expectation of γ and of $(1-\beta)$ both tend toward the same value – the average of the true power and prevalence – as n increases, rather than tending toward their respective ground truths. As we will illustrate in Example C, changing the significance level α can also change the binary outcomes of the within-subject NHSTs thus changing the observed k, which can lead to drastically different posterior estimates when n is small. Ideally, an estimator would not be so sensitive to the value of an arbitrary parameter.

The extended Binomial model can be used for between-group comparisons of prevalence or within-subject power in the same manner we described for the *p*-curve mixture model (see ***Between-group differences***), and we compared the sensitivity of that approach using the two models in our simulations. Ince et al. (2021) also describe an analog approach to estimating within-group differences, but we did not implement that version of their model here.

4. Bayesian model comparison

An advantage of framing the problem of estimating prevalence and effect size as inferring on the parameters of a generative model is that the model can be compared using the full suite of Bayesian model comparison tools. In particular, the null (all H_0) model, the alternative (all H_1), and the mixture (some subjects H_0 and others H_1) can all be compared, as all of these models are described by *p*-curve likelihoods. Critically, this allows a researcher to perform Bayesian hypothesis testing

to assess whether the population distribution of an effect is indeed heterogenous. Notably, since the prevalence is a proportion, the posterior HDI will never contain 0 nor 1; thus, the mere fact the HDI excludes these values cannot be used as evidence for population heterogeneity. Only model comparison can be used to support claims about the presence of absence of population heterogeneity when the HDI is close to these boundaries.

One approach to Bayesian model comparison is to estimate the leave-one-(subject)-out cross-validated posterior likelihood of the data, which can be computed – with an uncertainty estimate – efficiently using Pareto-smoothed importance sampling or PSIS (Vehtari et al., 2017, 2024). Intuitively, the model under which the likelihood of the held-out data is the highest is also the most likely to explain new data. "Weights" can be computed for each model which are, roughly speaking, a probability that each model will predict new observations better than the other models (Vehtari et al., 2017). Because this method of model comparison is both computationally efficient and relatively numerically stable, we have incorporated it into the *p2prev* package so that users can estimate leave-one-out likelihoods and model weights with just a couple of lines of code.

However, the PSIS approach's emphasis on prediction may not always be desirable. If, for example, the true prevalence of an effect is zero, a *p*-curve mixture model should estimate a posterior prevalence near zero, resulting in posterior predictions quite similar to those of the H_0-only model. PSIS may not differentiate between these edge cases as well as Bayes Factors, which quantify how much the prior odds on which model is *correct* (rather than most predictive) should be updated given the observed data. We estimate Bayes Factors by nesting the H_0-only, H_1-only, and H_0/H_1-mixture models within a single, hierarchical model with a Categorical prior over models (Kruschke, 2014). This nested model is numerically difficult to sample from due to the discrete latent variable, which can require troubleshooting the sampler; consequently, we provide a code example for estimating Bayes Factors in *PyMC* (corresponding to Example C) but do not build this method into the *p2prev* package itself as we cannot pick default sampler settings that will work in all cases. It should be noted that Bayes factors can be sensitive to the effect size (δ) prior, so researchers should ensure their choice of δ prior does not unduly affect their results if they deviate from validated default priors.

In cases where the population distribution is indeed heterogenous, the H_0/H_1-mixture model should be sufficiently distinguishable from the H_0-only and H_1-only models using the PSIS method. In cases where the sample size is very small or if one wishes to support a claim that either the H_0-only or the H_1-only model is true, we recommend using Bayes Factors instead as predictive criteria may not discriminate the models. In addition, for an approach that is agnostic to the effect size under H_1, researchers can also test the H_0-only model in the frequentist framework by combining *p*-values across subjects into a "global" *p*-value using Fisher's method, Simes method, or related approaches (Ganju & Ma, 2017).

5. Simulations

5.1. Prevalence estimation benchmark with classification accuracies

In this set of simulations, we aimed to benchmark the prevalence estimation performance of the *p*-curve mixture model and of the Binomial model across various sample sizes in a setting in which

within-subject statistical power to detect an effect using NHST was less than 100%. To this end, we simulated (1) a setting in which within-subject power was low (~0.6) but prevalence was high (~.95) and (2) a setting in which those were flipped and power was high (~.95%) and prevalence was low (~0.6). We performed 1,000 simulations at every sample size from $n = 3$ to $n = 60$, spaced by 3.

We also wanted to stress-test the *p*-curve mixture model by introducing some everyday violations of its assumptions. The way we have specified the *p*-curve mixture model is such that all subjects for which H_1 are modeled as having the same effect size δ. We intend this to be interpreted as the "average effect size given H_1," but we would like to verify that our specification can tolerate mild over-dispersion of effect size. Consequently, we simulated 50 trials from subjects with within-subject classification accuracies, for subjects in which H_1 was true, were drawn from a distribution with (1) mean 0.65 ± 0.01 SD in the low-power setting and (2) mean 0.75 ± 0.01 SD in the high-power setting. We computed *p*-values for each subject using a permutation test to show that the within-subject NHST used to generate *p*-values does not need to make normality assumptions or even be parametric, despite the presence of the normal density function in Equation 1. This choice is also motivated by the fact that it is common to evaluate classification performance using permutation tests when one has estimated out-of-sample accuracies by cross-validation, such as in multivoxel pattern analysis in neuroimaging where it is assumed the assumptions of the binomial test are violated (Valente et al., 2021).

On each simulation, we compute the posterior expectation of the population prevalence and the width of the 95% HDI (smaller means less uncertainty) under both the *p*-curve and Binomial models, and we compute the frequentist false coverage rate of the HDI at each sample size, which is the proportion of simulations in which the HDI contained the true prevalence. While Bayesian HDIs do not provide coverage rate guarantees like frequentist confidence intervals, it is nonetheless a desirable property when a 95% HDI achieves a false coverage rate near or less than 5% in practice, which can be evaluated by simulation (Gelman & Carlin, 2014).

5.2. Within- and between-group difference estimation with EEG data

In this set of simulations, we aimed to assess the sensitivity of the *p*-curve model for detecting between- and within-group differences in population prevalence and in within-subject effect size with realistically noisy data. We compare the *p*-curve model's sensitivity to that of group-level null hypothesis significance testing, though NHST does not differentiate between prevalence and within-subject effect size contributions to detected changes in the group mean effect size. In the between-group simulation, we also compared the model's ability to differentiate between changes in prevalence and effect size to that of the Binomial model.

We simulated electroencephalographic (EEG) event related potential (ERP) data by adapting the simulation code used by Sassenhagen and Draschkow, in which a simulated ERP is inserted into background noise that has been cut out of a real EEG recording (Sassenhagen & Draschkow, 2019). This approach ensures that the noise properties reflect that of realistic EEG data. On each simulation, we simulated 100 trials of EEG data (50 per condition) for each of 30 subjects per group. We assigned each subject to H_0 or H_1 with probability equal to the population prevalence $\gamma = 0.45$, and we inserted an ERP into one condition of H_1 subjects' data, which we

adjusted the size of until the power of a within-subject NHST (at significance level 0.05) was also approximately 0.45. As is the gold-standard in the EEG literature, we computed within-subject *p*-values with an independent-samples cluster-based permutation test which tests for an effect anywhere across all electrodes and timepoints while controlling the familywise error rate (Maris & Oostenveld, 2007). Naturally, the cluster-level test statistic in this NHST does not have a simple distribution under the null hypothesis – EEG data is subject to substantial noise that is autocorrelated across both electrodes and time – so this is meant to be a strong demonstration of the power of *p*-curve mixture models where it would be impractical to construct a parametric Bayesian mixture model.

In each between-subject simulation, we simulated another group of 30 subjects with either higher H_1 prevalence (0.45→0.95) or higher within-subject power (0.47→0.96), which was achieved by increasing the magnitude of the simulated ERP. For each within-subject simulation, we generated another 100 trials per subject with similarly increased power or prevalence for some test H_2, ensuring that H_2 is true for all subjects in which H_1 was true. This latter scenario reflects how most ERP studies would be designed; a 2x2 design in which one factor is designed to isolate an ERP component of interest (i.e. compute a difference wave), and another factor (i.e. an experimental manipulation) is meant to induce a change in that ERP component.

We perform 1,000 simulations of each of those four contrasts. We then quantify how sensitive the models' posterior probability of a prevalence increase across groups/conditions is at correctly identifying prevalence increases without false alarming on power increases and vice versa, using the area under the receiver-operator characteristic curve (AUROC) and the true positive rate at a 5% false positive rate. We also apply a group-level NHST on each simulation, first computing the subjects' difference waves (average of 50 trials in one condition minus the average of the 50 trials in the other condition) and inputting them into an independent samples cluster-based permutation test for the between-group simulations and into a paired cluster-based permutation test for the within-group simulations (Maris & Oostenveld, 2007).

6. Examples

We also describe examples that demonstrate the utility of the *p*-curve mixture modeling on real data. In Example A (see **Results)**, we reproduce the custom mixture model we used in a previous study in which we tested (cardiac) interoceptive accuracy (Veillette, Gao, et al., 2024), and we compare the resultant prevalence estimates to those obtained from a *p*-curve mixture model. In Example B, we apply *p*-curve mixture models to a published dataset on absolute pitch (AP) memory (Hedger et al., 2020), evaluating whether the effect of tonal language experience on pitch memory is dichotomous (i.e. tonal language affects whether one has AP) or grated (i.e. language experience effects pitch recognition accuracy, given that one already has AP). Lastly, Example C illustrates how meaningful prevalence inferences can be made – albeit with high uncertainty – from only a handful of densely sampled subjects, applying *p*-curves for estimation and model comparison in an increasingly common "precision neuroimaging" setting (Poldrack, 2017). In Example D, we estimate within-subject prevalence differences, and conditional prevalences of one hypothesis given another, for EEG decoding results – both for decodable information throughout the whole epoch and across time. We also estimate within-subject power.

7. Data and Code Availability

The raw datasets used in our examples are available at the open data repositories documented in the papers in which these data were originally reported (Hedger et al., 2020; Veillette, Chao, et al., 2024; Veillette et al., 2023; Veillette, Gao, et al., 2024). However, we also provide smaller, preprocessed versions of these datasets sufficient to replicate our examples in the same repository as our code. Our GitHub repository (https://github.com/john-veillette/p2prev) contains the source code for the *p2prev* package, tutorial examples including the examples in this paper, and code to reproduce our simulations. The GitHub repository will be continually updated while we make improvements to *p2prev* and its documentation, but all stable releases of the *p2prev* source code and accompanying tutorial code are permanently archived on *Zenodo,* with a digital object identifier (DOI) that always leads to the most recent release (https://zenodo.org/doi/10.5281/zenodo.11459064) as well as a unique DOI for each release (e.g. the version used for the simulations in this report, *p2prev* v0.0.2). The *p2prev* package can also be installed from the PyPI package server using Python's "pip" command.

Results

Example A: Interoception

In a previous study, we used a custom mixture model to estimate the population prevalence of those who can feel their own heart beating (Veillette, Gao, et al., 2024). Subjects saw two circles pulsing, side-by-side, on the screen in front of them; one circle was synchronized to their cardiac systole (the phase in which blood pressure increases following a heart contraction, triggering baroceptors in the arteries) and the other pulsed exactly anti-phase. Subjects were asked, on each trial, to guess which circle was synchronized to their heartbeat sensations.

As described in **Methods: Bayesian mixture models for prevalence estimation**, we originally modeled the number of correct trials k_i for each subject i as $k_i \sim \text{Binomial}(n_{\text{trials}}, 0.5)$ under H_0 and as $k_i \sim \text{Binomial}(n_{\text{trials}}, \kappa_i)$ for some accuracy $\kappa_i \in (0.5, 1.0)$ under H_1, and we sampled from the posterior of that mixture model given the observed k_i's from 54 subjects to obtain an estimate of the prevalence of H_1 in the sampled population (i.e. above-chance interoceptive accuracy). The posterior expectation for the prevalence of cardiac perceivers was 0.11 (95% HDI: [0.01, 0.21]).

Now, with *p*-curve mixture models, we can estimate the same quantity without constructing and programming a custom model. Converting the subjects' accuracies to *p*-values using a one-tailed binomial test and then plugging those *p*-values into the *p2prev* package yielded a nearly identical posterior (M = 0.10, 95% HDI: [0.01, 0.20], see Figure 2c). It is also possible to estimate the per-subject probability of H_1 in both models – this was, in fact, what we were actually trying to do in the original study – and we find even these per-subject probabilities are almost the same in the two models (see Figure 2d). Still, it is worth noting one disadvantage of the *p*-curve model: the custom mixture model yields an effect size estimate in units of accuracy, while the *p*-curve model only gives an uninterpretable, unitless effect size (although that effect size parameter can be converted into a statistical power at a chosen significant level).

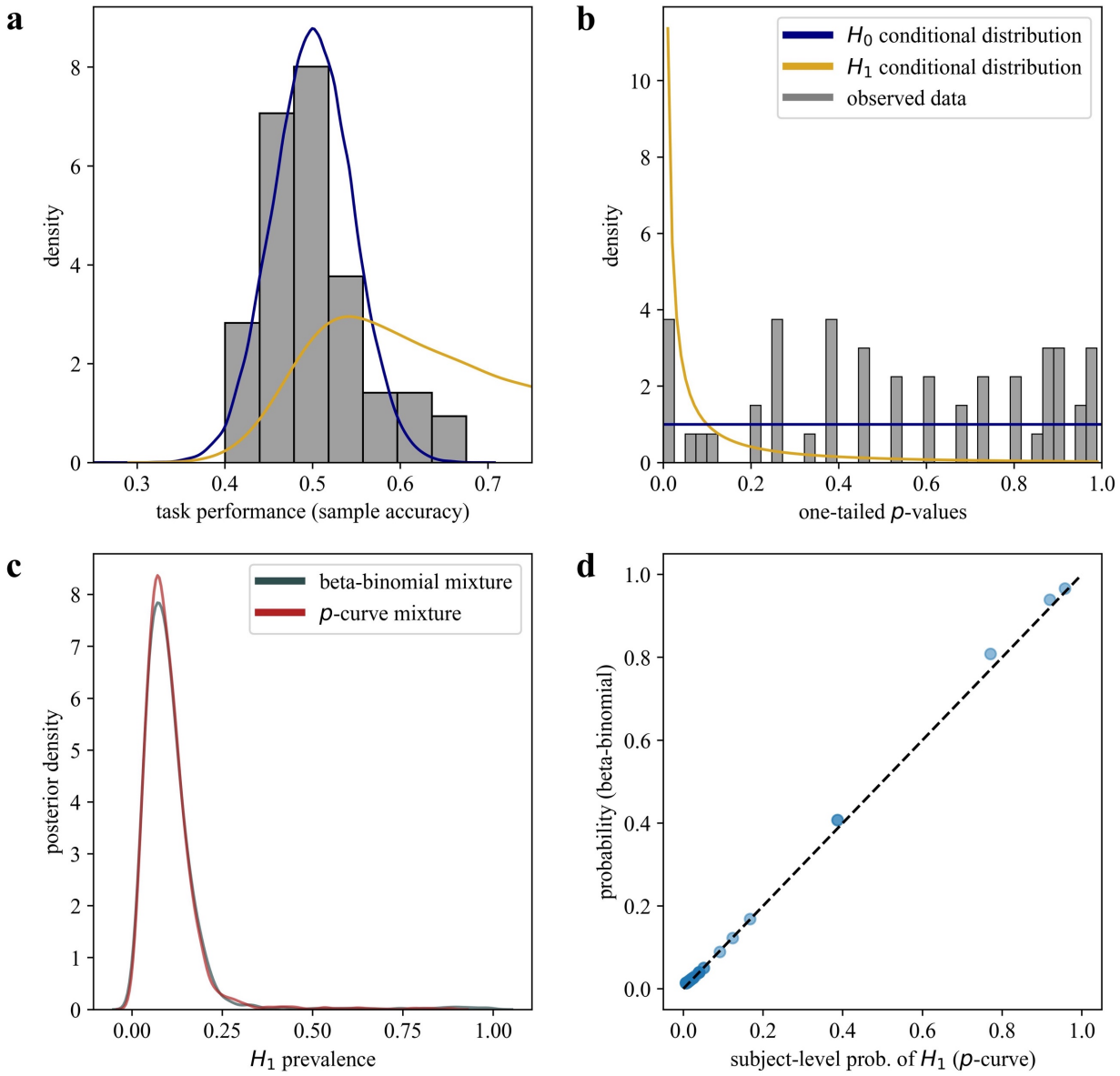

Figure 2: Null hypothesis significance tests as transformations of variables. (a) In a previous study, we estimated the prevalence of above-chance performers on a discrimination task by modeling the distribution of subjects' accuracies as a mixture between a Binomial distribution for subjects for whom H_0 was true (at-chance accuracy) and a Beta-Binomial distribution for subjects in which H_1 was true (above-chance accuracy). (b) We could have instead converted the accuracies to p-values using a binomial test and modeled the distribution of p-values as a mixture between two p-curves. (c) Both models result in almost identical posterior distributions for the population prevalence and (c) identical per-subject posterior probabilities of H_1.

Prevalence estimation benchmark

Like essentially all Bayesian estimators, the p-curve mixture model's posterior mean is a biased estimator; it is heavily influence by the posterior distribution when the sample size is small but gets closer to the true prevalence as n increases (see Figure 3a). The Binomial prevalence model (Ince et al., 2021), however, is unable to distinguish between population prevalence and within-

subject power, so its posterior mean converges near the average of the two as the number of subjects increases.

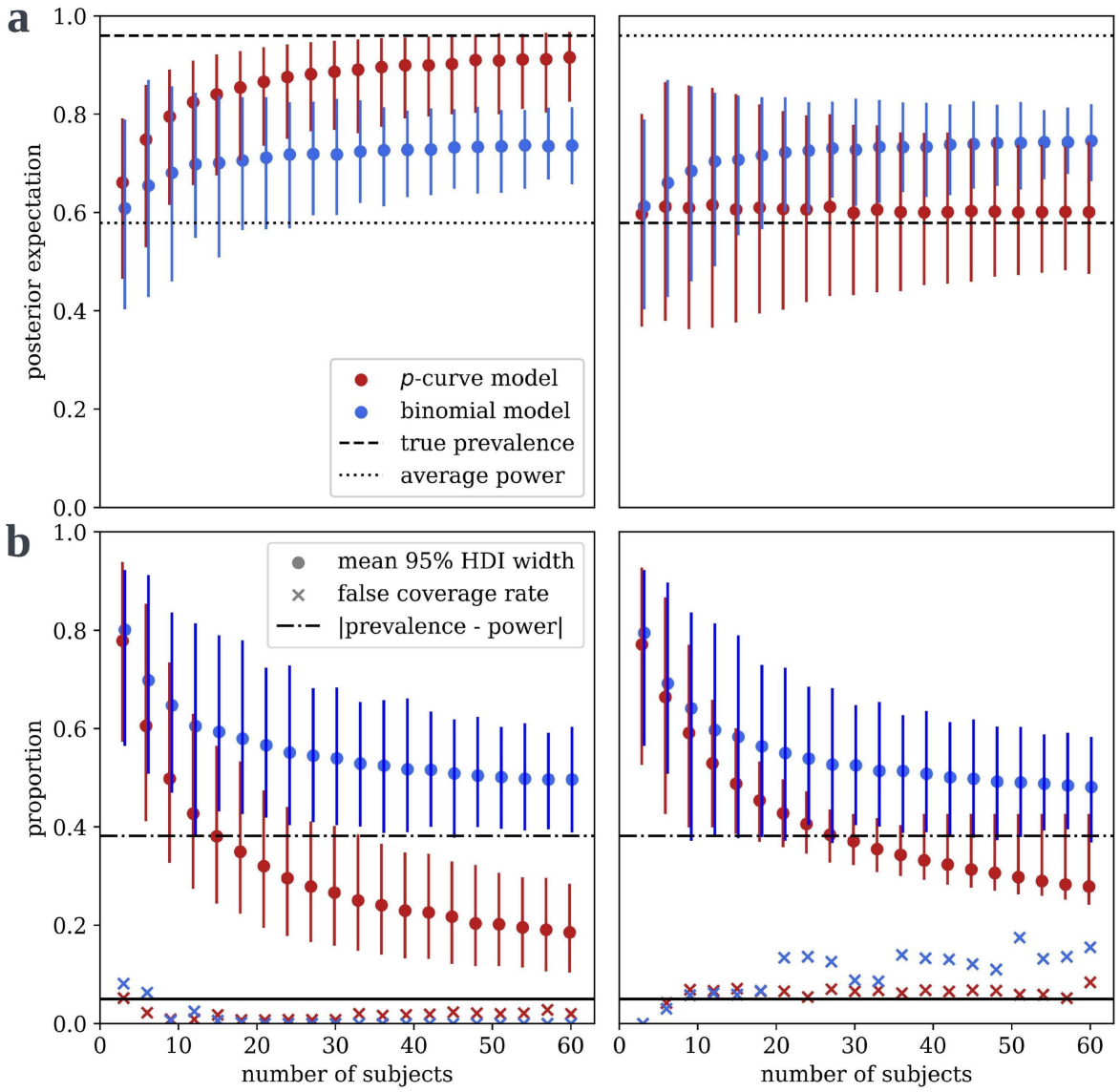

Figure 3: *p*-curve mixture model outperforms state-of-the-art when within-subject power is not known. In all panels, dots reflect the mean across all simulations and bars contain the values from 95% of simulations. (a) The *p*-curve mixture model's expected population prevalence (a.k.a. posterior mean) gets closer to the true prevalence as it observes more data, but the Binomial model cannot differentiate between prevalence and within-subject power, so its posterior mean converges somewhere between the two. (b) The *p*-curve's highest density interval (HDI), which contains 95% of posterior probability, shrinks as the model is given more data, reflecting greater certainty, but the Binomial model's HDI width is lower-bounded by the difference between power and prevalence. The *p*-curve mixture model maintains a false coverage rate (proportion of simulations in which the true prevalence is not contained in the HDI) comparable to that of a frequentist confidence interval.

The *p*-curve mixture model gets (justifiably) less uncertain when it has more data; that is, the interval that contains 95% of the posterior probability (95% HDI) shrinks as the number of

subjects gets larger (see Figure 3b). Conversely, as the Binomial model cannot distinguish between high-prevalence/low-power and low-prevalence/high-power, its HDI is almost never smaller than the difference between power and prevalence. Thus, after some time, additional data ceases to be useful to the Binomial model, but the *p*-curve mixture model continues to learn from new data. Notably, the *p*-curve mixture model, though it does not come with a false-coverage rate guarantee like a frequentist method would, yields 95% HDIs that achieve empirically strong false coverage rates near or below 5% much like a frequentist 95% confidence interval.

Sensitivity to between- and within-group differences in EEG data

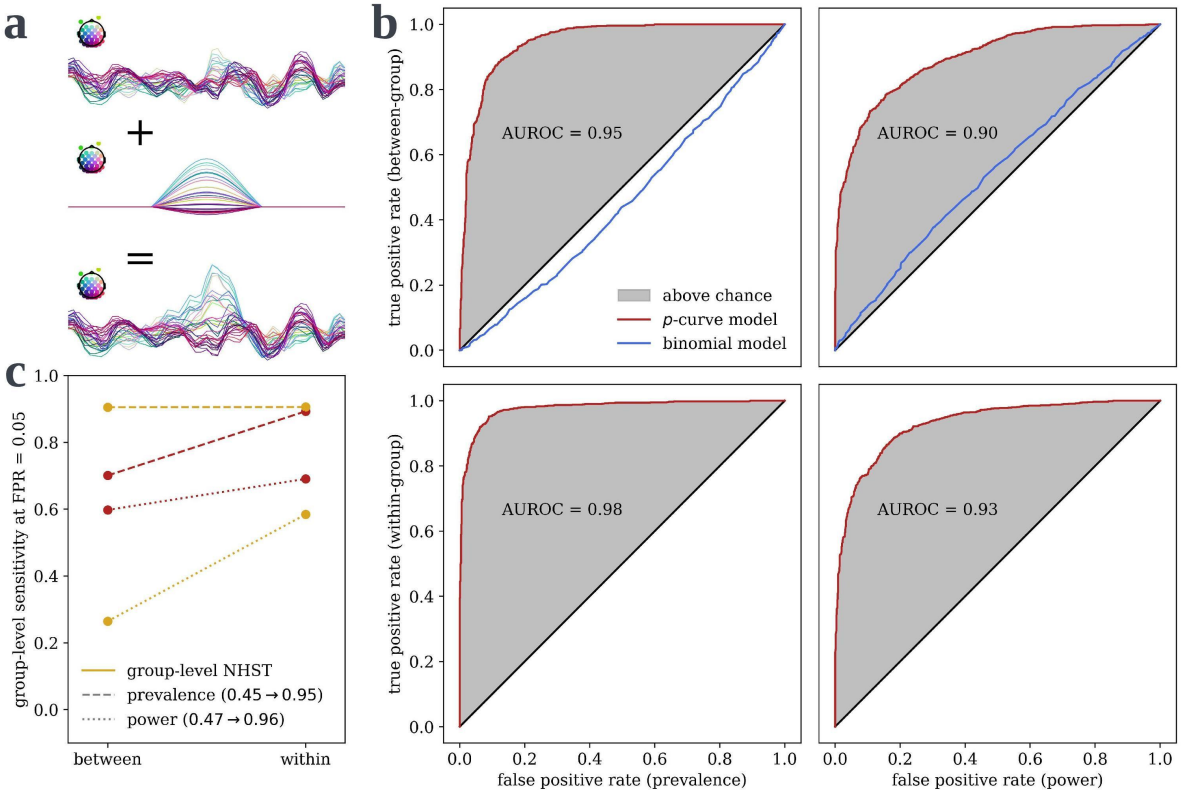

Figure 4: *p*-curve mixture models can detect and discriminate between differences in effect prevalence and effect size between groups or conditions. (a) We simulated EEG data by inserting a simulated evoked response into background noise from a real EEG recording. (b) On each simulation, we simulated two groups of *p*-values for within-subject tests of the evoked response, manipulating either the prevalence of the evoked response, on left, or its magnitude in those who show the effect, on right. The model was highly sensitive at detecting prevalence or effect size differences between independent groups of subjects, on top, or between two within-subject conditions, on bottom. False positives, here, refers to mistaking a prevalence increase for a power/effect size increase or vice versa. (c) The detection rate at the 5% false positive rate is compared to the sensitivity of a group-level NHST with significance level 0.05. In contrast to the normative interpretation of a significant difference in group mean, NHST was highly sensitive to changes in prevalence but less sensitive to changes in effect size than *p*-curve mixtures. In the within-group case, there is no apparent sensitivity cost to using *p*-curve mixtures, which can dissociate between differences in prevalence and within-subject power/effect size.

The *p*-curve mixture model was highly sensitive to changes in population prevalence and in within-subject effect size in our simulation (see Figure 4). The model is sensitive to between- or within-

group differences in prevalence or in within-subject power without mistaking one for the other (see Figure 4b). Moreover, while significant differences in the group mean effect size are often interpreted as differences in within-subject effect size, NHST was highly sensitive to changes in the population prevalence (see Figure 4c). Actually, *p*-curve models outperformed NHST at detecting changes in within-subject effect size, likely because the *p*-curve model's effect size estimate isolates those subjects who actually show the effect – implicitly serving the function of outlier removal but without imposing an arbitrary threshold.

It is worth noting that, while the *p*-curve mixture models obtained strong sensitivity at a 5% false positive rate (see Figure 4c), the cutoff at which this 5% rate was obtained – that is, the posterior probability of a prevalence or power increase at or above which one would say that there is, in fact, an increase in prevalence or power – is not necessarily 95%. In our simulations, the posterior probability threshold that yielded a 5% false positive rate for differences in effect size was roughly 95% (0.958 and 0.962 for between- and within-groups respectively), but was much lower for differences in prevalence (0.846 and 0.752 respectively). This is not a bad thing; actually, it is a hallmark of high specificity. Our posterior probability of a prevalence difference is insensitive to changes in within-subject effect size, which is what we want. The true Bayesian approach would be to simply report to posterior probabilities, but pragmatically some researchers may indeed have a specific reason to care about frequentist false positive rates. Such a researcher could always take a brute-force approach – at the cost of computation – and use the *p*-curve model's posterior probability as the test statistic for a permutation test, and thus obtain a frequentist *p*-value for prevalence or effect size differences, though we think this is not usually necessary.

Example B: Pitch Perception

Absolute pitch (AP) – the ability to recognize musical notes by their pitch alone -- is generally cited as quite rare with only 1 in 10,000 demonstrating this; even trained musicians normally require the aid of a reference note (Takeuchi & Hulse, 1993). A common claim in the literature is that experience speaking a tonal language increases the likelihood that one will develop absolute pitch. However, studies tend to dichotomize subjects into AP and non-AP groups based on whether they exceed some threshold accuracy. The number of subjects who exceed any binary threshold, regardless of whether it was set arbitrarily or by some statistical criteria, is in principle a function not just of the prevalence of AP but also of the pitch-labelling accuracy of those subjects who do have AP. Thus, an alternative explanation for a higher proportion of tonal language speakers clearing some threshold is an increase in effect size, not in prevalence.

In a previous study (Hedger et al., 2020), we collected behavioral judgments in a pitch-labeling task from a large sample of online participants – many of whom were attracted, unconventionally, by an article in the Wall Street Journal (Mitchell, 2017). Here, we analyze that dataset to estimate the difference in AP prevalence and in within-subject effect size. We first input the one-tailed *p*-values given by a binomial test on subjects' accuracies into *p2prev* to fit a *p*-curve mixture model on the full sample, estimating a prevalence of 0.53 (95% HDI: [0.46, 0.61]). It is important to note that prevalence estimates are always for the *sampled population,* which is usually not the general population. Here, our subjects opted to take an online quiz to see if they have AP after reading about it in a newspaper article, so our prevalence estimates likely refer to a population that has self-selected for believing they are above-chance at naming musical notes.

Then, we fit *p*-curve mixture models to tonal language speakers and other subjects separately, and we subtract posterior samples between groups to approximate the posterior of the difference (see **Methods 2.5**). This results in a 95% highest density interval of [-0.02, 0.35] for the prevalence increase as a result of speaking a tonal language – not evidence against a prevalence increase by any means but the HDI does still contain zero as a plausible difference. Interestingly, however, the HDI for the within-subject effect size substantially departs from zero (95% HDI: [0.18, 1.22]), which is fairly strong evidence that tonal language experience increases the within-subject effect size *given that a subject already has AP*, seemingly in contrast to how the effect of tonal language is framed in the literature.

Of course, the dataset on which we did this analysis is quite idiosyncratic, so we do not mean to suggest that the AP literature should reevaluate its canon based merely on Example B in a methods paper. However, while our **Introduction** and simulation results warn against interpreting differences in the mean effect size as effect size differences per se, this example nicely illustrates that putative prevalence differences may also turn out to be accounted for by effect size differences when subjected to additional statistical scrutiny. Theoretical assumptions should always be evaluated explicitly, and *p*-curve mixture models provide a broadly applicable tool for doing so.

Example C: Precision fMRI

It is becoming increasingly common in the neuroimaging literature to collect lots of data from very few subjects, rather than a bit of data from many subjects as in a traditional group study. This "precision neuroscience" approach allows one to detect within-subject effects that would wash out in a group average due to poor spatiotemporal alignment across subjects or other idiosyncrasies in the functional organization of the brain (Poldrack, 2017). However, as such studies tend to forgo group-mean inference in favor of within-subject statistics, the extent to which results should be expected to generalize to the population is usually left to be inferred by the reader. While some researchers have suggested effects strong enough to be observed in a single subject should be assumed to be nearly universal (Dosenbach & Gordon, 2023), this inference is contradicted by the empirical observation that large effect sizes tend to be associated with *more* rather than less heterogeneity (Olsson-Collentine et al., 2020). Such generalization claims would be substantiated empirically if explicitly support using prevalence statistics and inference. Indeed, since these densely sampled studies already perform significance testing within each subject – just as required for a *p*-curve mixture model – some researchers have already proposed that prevalence statistics are the best way to combine results across subjects (Ince et al., 2022).

In a recent study, we collected 100 minutes of fMRI data (as part of a longer experiment) from 4 subjects while they performed a motor task in an MRI scanner and we recorded their hand movements via motion tracking (Veillette, Chao, et al., 2024). We attempted to predict subjects' continuous brain activity from the internal representations of a computational model that performs the same motor task in a biomechanical simulation, aiming to approximate the "inverse kinematic" computation required to generate muscle movement activations that will move the hand to a target position. We obtained out-of-sample R^2 values from this theoretically-motivated model and from a control model, and we compared R^2's non-parametrically using threshold-free cluster

enhancement to obtain a familywise error rate (FWER) corrected *p*-value for every voxel in cortex (Smith & Nichols, 2009). As suggested by Ince and colleagues, the lowest FWER-corrected *p*-value across all voxels can be used as a "global" *p*-value for the presence of an effect anywhere in the brain (Ince et al., 2021). Similarly, the lowest FWER-corrected *p*-value in a region of interest (with FWER correction across the whole brain, not just that ROI) is a *p*-value for the presence of an effect in that region. This allows researchers to abstract over spatial misalignment when aggregating results across subjects.

The smallest FWER-corrected p-values across cortex for each subject were 0.00060, 0.02999, 0.04939, and 0.94601, so we could reject the null hypotheses that our theoretical model does not outperform the control model in ¾ subjects at significance level $\alpha = 0.05$ or in ¼ subjects at significance level $\alpha = 0.01$. Consequently, when we applied the Binomial prevalence model, we obtained totally different estimates depending on whether we used $\alpha = 0.05$ (prevalence = 0.721, 95% HDI: [0.382, 1.00]) or $\alpha = 0.01$ (prevalence = 0.504, 95% HDI: [0.114, 0.993]). Our discomfort with this estimator's dependence on an arbitrary parameter is what motivated us to develop *p*-curve mixture models in the first place. When we put our *p*-values into *p2prev,* we estimated a prevalence of 0.610 (95% HDI: [0.224, 0.972]). As seen in Figure 3, the 0.610 posterior expectation is likely not very meaningful at such a small sample size; however, the HDI bounds maintain strong false coverage rate properties even at very low sample sizes and are thus informative.

While three of our subjects had low within-subject *p*-values, one subject had a much higher *p*-value. Is this enough evidence to suggest that the subject showed no effect, or is it the case that we were just underpowered to detect it? Do we even have enough evidence to support that our subjects with low *p*-values do show an effect or were they statistical flukes? *P*-curve models can help with this as well, as we can compare our mixture model to *p*-curves for the null and alternative hypothesis alone. When we calculated Bayes Factors (see **Methods 4: Bayesian Model Comparison**), we obtained BF = 42.03 when comparing the mixture model vs. the H_0-only model, indicating that the observed data were more than 42 times more likely under the mixture model in which the null hypothesis is always true. This is strong evidence that our subjects indeed show an effect. When we compare the mixture model to the H_1-only model, on the other hand, we obtained a much lower Bayes Factor of BF = 3.86. While evidence leans in favor of the mixture model, indicating that a model in which not all subjects express the effect can explain the high *p*-value better than the H_1-only model, the evidence is not resounding. (For reference, some journals require Bayes Factors of at least 10 to support claims, though this is somewhat arbitrary). Of course, evidence should not be resounding; we only saw one low *p*-value! Nevertheless, we obtain informative results even with a small sample size, and Bayesian model comparison provided a rigorous way to collectively evaluate our within-subject results without requiring us to use an arbitrary α threshold.

Example D: EEG Decoding

In a recent study with 25 subjects, we used functional electrical stimulation (FES) of arm muscles to usurp subjects' intentional motor control in a response time task (Veillette et al., 2023). By carefully timing the latency of FES to preempt subjects' volitional movements, we were able to elicit electrically-actuated finger movements that, under controlled timing conditions, subjects

either claim they themselves caused the movement or that they did not cause the movement. We accomplished this using an adaptive procedure that, for each subject, estimated and simulated at the FES latency at which subjects responded – when asked after each trial whether they or the FES caused the button press in the reaction time task. The timing was set so that that subjects reported that they caused the movement on roughly 50% of trials. This approach created balanced sets of trials subjects perceived as self- or other-caused, so that their sense of agency (SoA) could be decoded from their EEG.

While we reported group-averaged results in the original study (Veillette et al., 2023), as is standard, in analysis of the data we noted substantial heterogeneity in how subjects responded to the adaptive procedure. In some subjects, the adaptive procedure honed in on their threshold early and FES remained at that threshold latency for the rest of the experiment. In these subjects, a within-subject logistic regression predicting agency judgments from FES latency would return significant results, as subjects were highly sensitive to deviations from their threshold. However, other subjects' thresholds seemed to slowly shift over time throughout the experiment; in these subjects, the same logistic regression would yield non-significant results, as the same 50/50 response distribution was observed across a range of latencies. With *p*-curve mixture models, we can now better assess how decoding of agency judgments from EEG is related to this behavioral difference.

Using a within-group prevalence model, we can model the joint probability of subjects' agency judgments being sensitive to FES latency around their threshold (i.e. are "stable-threshold subjects" vs. "unstable-threshold subjects") and of agency judgments being decodable from their EEG – for which we computed a *p*-value for the 10-fold cross-validated decoding AUROC using a permutation test at each time-point in the epoch following FES onset. As described by Ince et al. (2021), we can perform prevalence inference at each time in the EEG epoch, or we can take the lowest familywise error rate corrected *p*-value across subjects as a "global" *p*-value testing the null hypothesis that agency judgments are not decodable from their EEG at any point. Familywise error rates corrected p-values were compute using the maximum statistic method (Nichols & Holmes, 2002). Using the lowest familywise error rate corrected *p*-value in this way satisfies the assumptions of the *p*-curve mixture model, as its null distribution is uniform, but this would not necessarily be the case for, say, a false discovery rate corrected *p*-value.

Using the within-group *p*-curve model described in **Methods**, we estimate the prevalence of behavioral sensitivity of agency judgments to FES latency – over the whole epoch – as 0.655 (95% HDI: [0.473, 0.870]) and that of decodable agency judgments as 0.922 (95% HDI: [0.831, 0.996]), with evidence for a prevalence difference of 0.268 (95% HDI: [0.054, 0.479]). Logically, as the prevalence of decodable agency judgments exceeds that of the behavioral sensitivity effect, we can already conclude that above-chance decoding performance can be achieved for both behavioral phenotypes. However, our within-group prevalence model also allows these conditional probabilities to be estimated explicitly. The prevalence of EEG-decodable agency judgments among stable-threshold subjects is estimated as 0.942 (95% HDI: [0.835, 1.000]) and among unstable-threshold subjects is 0.881 (95% HDI: [0.668, 1.000]), without strong evidence for a difference (0.061, 95% HDI [-0.169, 0.321]). However, just because there is not a difference in prevalence *across the whole EEG epoch* does not mean there is no difference at all. We can also calculate the conditional prevalences across time as seen in Figure 5.

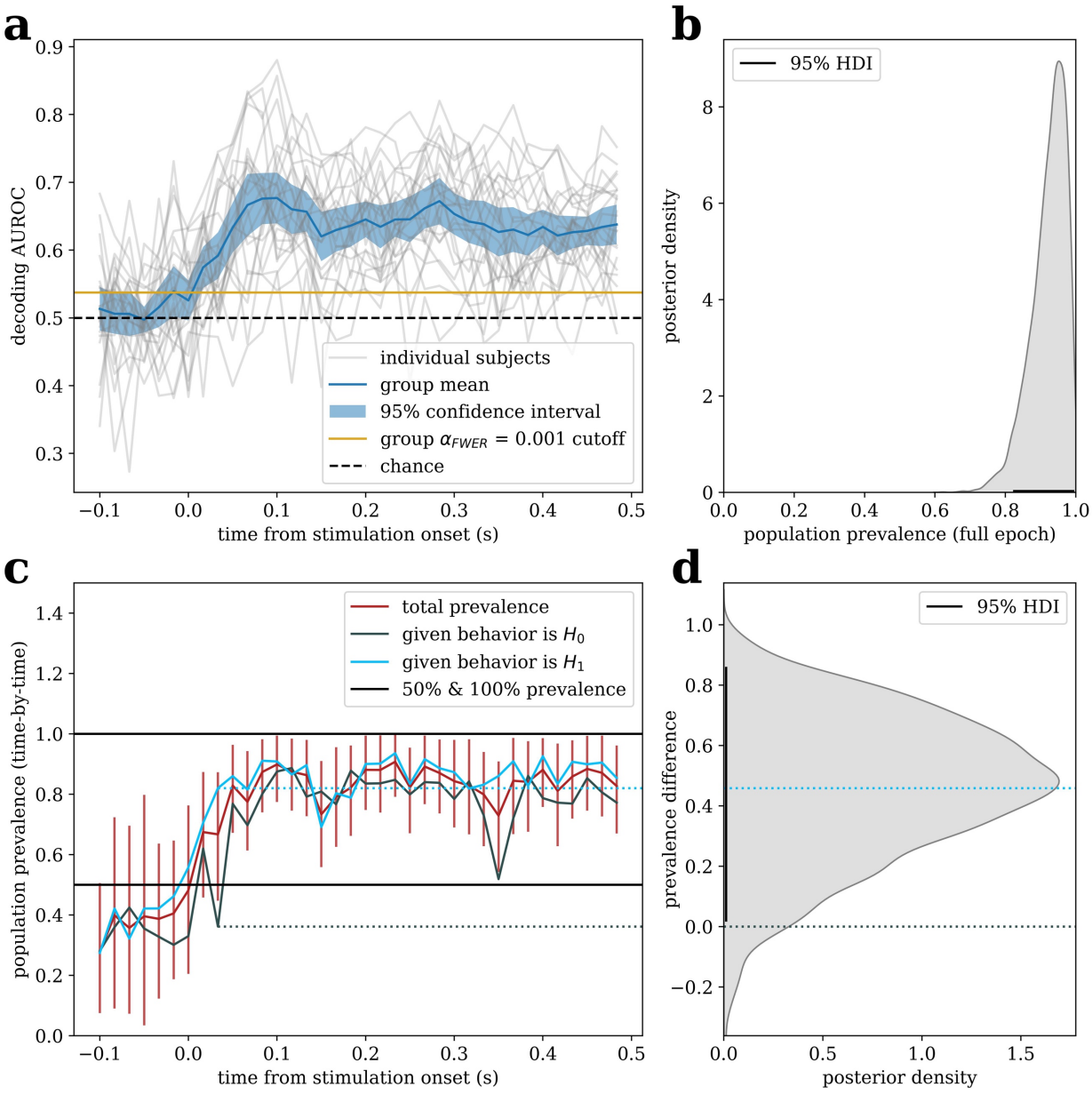

Figure 5: Prevalence estimation across time with EEG decoding results. (a) The group-mean and single subject decoding performance time courses for predict agency judgments from the EEG as described in Example D. (b) The posterior for the population prevalence of decodable agency judgments at any point in the epoch. (c) The prevalences of decodable agency judgments at specific times in the EEG relative to electrical muscle stimulation, with posterior means for both total prevalences and prevalences conditional on the behavioral effect, and 95% HDIs for total prevalence. (d) The posterior prevalence difference conditional on the behavioral effect at 0.033 seconds, showing stable-threshold subjects have a higher prevalence of decodable information at this time.

As seen in Figure 5c, the prevalences conditional on being a stable-threshold or unstable-threshold subject can be computed at each time following stimulation, and a posterior difference can be estimated. For example, at 0.033 seconds after stimulation, just after the initial cortical response – which is expected at around 20 milliseconds for stimulation on the wrist (Anziska &

Cracco, 1981) – decodable information is already present in a majority of stable-threshold subjects (0.820, 95% HDI: [0.588, 1.000]). However, unstable-threshold subjects are less likely to show decodable information at this time point (prevalence = 0.361, 95% HDI: [0.022, 0.688]; difference = 0.458, 95% HDI: [0.016, 0.861]). In other words, the neural responses to muscle stimulation tend to predict agency judgments earlier in stable-threshold subjects, speculatively reflecting increased sensitivity to low-level sensorimotor mismatches.

Notably, while it is possible that bifurcating subjects based on their judgment stability might allow group-mean approaches to detect a behavior-contingent difference in *mean* decoding performance, the prevalence estimation approach is able to support the claim that decodable information is *absent* at this time in a higher proportion of unstable-threshold subjects. As information-based measurement such as decoding accuracies are known to show significantly above-chance group-mean performance even when decoding is only achievable in a minority of subjects (Allefeld et al., 2016), it is unlikely group-mean approaches – even those that can support null results in principle, such as equivalence tests or Bayes factors – could ever support such a finding unless prevalence were actually zero. Indeed, for this same reason, significant group-mean decoding results actually do not generally support the claim that decoding is possible in a majority of subjects; formal prevalence inference is required to assess the population generalization of any "multivariate pattern analysis" (MVPA) study, though this fact is frequently ignored in the literature (Allefeld et al., 2016; Hirose, 2021). This can be seen in Figure 5, however, as significant group-mean decoding performance precedes the time at which we can conclude that prevalence exceeds 50% of the population. Conversely, the group-averaged time courses in Figure 5a may give the impression that decoding time courses are smooth, but there are numerous instances where we can conclude the prevalence of decodable information is well below 100% in Figure 5c. This indicated that individual decoding time courses are heterogenous, which is not at all obvious from looking at the single subject time courses, which could easily be dismissed as noise around a population mean.

Finally, as out-of-sample prediction in common cross-validation schemes are not identically and independently distributed due to dependence across folds (Dietterich, 1998), and thus decoding performance metrics tend not to follow a known distribution, performing power analyses for MVPA studies is challenging even among other neuroimaging studies. As *p*-curve effect size estimates, though uninterpretable in-and-of-themselves, can be easily converted into within-subject power estimates at any false positive rate (see **Methods: 2.1**), we can report within-subject power estimates *post hoc,* which can be extremely useful for planning future research. In this example, the within-subject power for detecting above-chance decoding at any point over the EEG epoch at significance level 0.05 is 0.956 (95% HDI: [0.918, 0.988]). Note that, in addition to providing a valid point estimate, our approach yields a credible interval (or full posterior) for the within-subject power; in contrast, *post hoc* power estimates for group-mean NHSTs are essentially useless, providing no information beyond that given by the *p*-value itself (Althouse, 2021).

Discussion

When we began developing *p*-curve mixture models, we were anticipating they would be applied primarily to dense sampling studies, offering only modest benefits over prevalence estimation

methods that require the power of within-subject tests to be near 100% (Ince et al., 2021). It is important to note, however, that we obtain strong performance even when (simulated) subjects have fairly low trial counts; 50 trials per condition in an EEG as experiment, as in our simulations, is certainly quite modest. Moreover, we were surprised to learn that *p*-curve mixture models can, in some cases, be applied without any loss in sensitivity compared to group-level null hypothesis significance testing, as in our within-group difference simulations where we even saw sensitivity *gains* (see Figure 4c). Applying these models instead of, or in addition to, traditional NHST can provide a more complete description of effect distributions within and across populations – a critical tool, we believe, for a time in which population heterogeneity in both psychological traits and functional brain organization are increasingly discussed (Henrich et al., 2010; Poldrack, 2017).

Indeed, a strong limitation of existing methods for prevalence estimation has been that they only provide lower bounds on the population prevalence (Allefeld et al., 2016; Ince et al., 2021), or otherwise require neraly 100% statistical power to yield accurate estimates of the true prevalence (Ince et al., 2021). As such, prevalence estimation in the behavioral sciences has been primarily geared toward studies explicitly designed to achieve very high within-subject power (Ince et al., 2022). Our novel approach provides an estimate of prevalence regardless of the within-subject power of a study and can thus be applied to datasets from experiments originally designed to test for differences in group means. This wide applicability allows researchers to easily quantify, with appropriate uncertainty, the proportion of the population to which their findings are expected to represent. This metric may provide crucial insight into how well-suited basic science findings are for translation to clinical settings, and empirical prevelance measurements could shed light on the causes of non-replications as researchers debate the role of generalizablity in precipitating the replication crisis (Bolger et al., 2019; Botella et al., 2019; Grandy et al., 2017; Moreau & Corballis, 2019; Olsson-Collentine et al., 2020).

It should be kept in mind, however, that prevalence estimates only pertain to the population from which the study sample was drawn; metaanlaytic findings seem to suggest that some fields of behavioral science – such as brain training, video gaming, mindset, and stereotype threat research – are particularly prone to produce treatment effect distributions with multiple modes that would not be well captured by a *p*-curve mixture model (Moreau, 2021). In this vein, while quantifying prevalence is an easy practice for empirical researchers to adopt in the interest of improving the precision and scope of their claims, it may only be the first step required to assess generalization in a research program aiming to design meaningful interventions or influence public policy (Bryan et al., 2021).

Many subfields in the behavioral and biological sciences carry unchecked assumptions about what their group mean differences really, well, *mean*. By providing a user-friendly interface to *p*-curve mixture models with our *p2prev* software package, we hope to facilitate the use of Bayesian mixture models so that scientists can rigorously evaluate these assumptions. Doing so need not necessarily entail collecting new data or adjusting experimental designs; our method is well-suited to glean novel insights from existing datasets – as illustrated by some of the worked examples in this paper. As such, we hope *p*-curve mixture models work their way into the experimentalist's toolkit.

References

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level *t*-test to prevalence inference. *NeuroImage*, *141*, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040

Althouse, A. D. (2021). Post Hoc Power: Not Empowering, Just Misleading. *Journal of Surgical Research*, *259*, A3–A6. https://doi.org/10.1016/j.jss.2019.10.049

Anziska, B., & Cracco, R. (1981). Short latency SEPs to median nerve stimulation: Comparison of recording methods and origin of components. *Electroencephalography and Clinical Neurophysiology*, *52*(6), 531–539. https://doi.org/10.1016/0013-4694(81)91428-0

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. S. (2010). mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software*, *32*, 1–29. https://doi.org/10.18637/jss.v032.i06

Bolger, N., Zee, K. S., Rossignac-Milon, M., & Hassin, R. R. (2019). Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology: General*, *148*(4), 601–618. https://doi.org/10.1037/xge0000558

Botella, J., Privado, J., Suero, M., Colom, R., & Juola, J. F. (2019). Group analyses can hide heterogeneity effects when searching for a general model: Evidence based on a conflict monitoring task. *Acta Psychologica*, *193*, 171–179. https://doi.org/10.1016/j.actpsy.2018.11.015

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, *10*(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Donhauser, P. W., Florin, E., & Baillet, S. (2018). Imaging of neural oscillations with embedded inferential and group prevalence statistics. *PLOS Computational Biology*, *14*(2), e1005990. https://doi.org/10.1371/journal.pcbi.1005990

Dosenbach, N. U., & Gordon, E. M. (2023). *Open Review of "Open Review of 'A somato-cognitive action network alternates with effector regions in motor cortex' (Gordon et al., 2023)" (Muret et al., 2023)*.

Frischkorn, G. T., & Popov, V. (2023). A tutorial for estimating mixture models for visual working memory tasks in brms: Introducing the Bayesian Measurement Modeling (bmm) package for R. *PsyArXiv Preprints*, *umt57*, Article umt57. https://doi.org/10.31234/osf.io/umt57

Ganju, J., & Ma, G. (2017). The potential for increased power from combining P-values testing the same hypothesis. *Statistical Methods in Medical Research*, *26*(1), 64–74. https://doi.org/10.1177/0962280214538016

Gelman, A., & Carlin, J. (2014). Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Grandy, T. H., Lindenberger, U., & Werkle-Bergner, M. (2017). *When Group Means Fail: Can One Size Fit All?* (p. 126490). bioRxiv. https://doi.org/10.1101/126490

Hedger, S. C. V., Veillette, J., Heald, S. L. M., & Nusbaum, H. C. (2020). Revisiting discrete versus continuous models of human behavior: The case of absolute pitch. *PLOS ONE*, *15*(12), e0244308. https://doi.org/10.1371/journal.pone.0244308

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hirose, S. (2021). Valid and powerful second-level group statistics for decoding accuracy: Information prevalence inference using the *i*-th order statistic (*i*-test). *NeuroImage*, *242*, 118456. https://doi.org/10.1016/j.neuroimage.2021.118456

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Ince, R. A., Kay, J. W., & Schyns, P. G. (2022). Within-participant statistics for cognitive science. *Trends in Cognitive Sciences*, *26*(8), 626–630. https://doi.org/10.1016/j.tics.2022.05.008

Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife*, *10*, e62461. https://doi.org/10.7554/eLife.62461

Kim, M., Van Horn, M. L., Jaki, T., Vermunt, J., Feaster, D., Lichstein, K. L., Taylor, D. J., Riedel, B. W., & Bush, A. J. (2020). Repeated measures regression mixture models. *Behavior Research Methods*, *52*(2), 591–606. https://doi.org/10.3758/s13428-019-01257-7

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mitchell, H. (2017, June 11). Can Perfect Pitch Be Learned? *Wall Street Journal*. https://www.wsj.com/articles/can-perfect-pitch-be-learned-1497178800

Moreau, D. (2021). Shifting Minds: A Quantitative Reappraisal of Cognitive-Intervention Research. *Perspectives on Psychological Science*, *16*(1), 148–160. https://doi.org/10.1177/1745691620950696

Moreau, D., & Corballis, M. C. (2019). When averaging goes wrong: The case for mixture model estimation in psychological science. *Journal of Experimental Psychology: General*, *148*(9), 1615–1627. https://doi.org/10.1037/xge0000504

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*(2), 101–122. https://doi.org/10.1016/j.jmp.2005.11.006

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. https://doi.org/10.1002/hbm.1058

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Patil, A., Huard, D., & Fonnesbeck, C. J. (2010). PyMC: Bayesian Stochastic Modelling in Python. *Journal of Statistical Software*, *35*(4), 1–81.

Poldrack, R. A. (2017). Precision neuroscience: Dense sampling of individual brains. *Neuron*, *95*(4), 727–729.

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*(6), e13335. https://doi.org/10.1111/psyp.13335

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). p-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). *Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015).*

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98. https://doi.org/10.1016/j.neuroimage.2008.03.061

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin*, *113*(2), 345.

Valente, G., Castellanos, A. L., Hausfeld, L., De Martino, F., & Formisano, E. (2021). Cross-validation and permutations in MVPA: Validity of permutation strategies and power of cross-validation schemes. *NeuroImage*, *238*, 118145. https://doi.org/10.1016/j.neuroimage.2021.118145

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., Gelman, A., Yao, Y., & Gabry, J. (2024). *Pareto Smoothed Importance Sampling* (arXiv:1507.02646). arXiv. https://doi.org/10.48550/arXiv.1507.02646

Veillette, J. P., Chao, A. F., Nith, R., Lopes, P., & Nusbaum, H. C. (2024). *Overlapping Cortical Substrate of Biomechanical Control and Subjective Agency* (p. 2024.07.24.604976). bioRxiv. https://doi.org/10.1101/2024.07.24.604976

Veillette, J. P., Gao, F., & Nusbaum, H. C. (2024). Cardiac afferent signals can facilitate visual dominance in binocular rivalry. *eLife*, *13*. https://doi.org/10.7554/eLife.95599.1

Veillette, J. P., Lopes, P., & Nusbaum, H. C. (2023). Temporal Dynamics of Brain Activity Predicting Sense of Agency over Muscle Movements. *Journal of Neuroscience*, *43*(46), 7842–7852. https://doi.org/10.1523/JNEUROSCI.1116-23.2023