Designing Experiments to Distinguish Theories*

Baiyun Jing Cassidy Shubatt

November 1, 2026

Click here for latest version

Abstract

We develop a measure of an experimental design's ability to distinguish between competing parametric theories. A design separates a theory from its competitors when these competitors are unable to reproduce behavior predicted by the theory. We propose an algorithmic procedure for constructing maximally-separating experimental designs, given a domain and hypothesized parameter values. We provide evidence that algorithmic designs improve upon the state of the art for design development and demonstrate how design separation can change the interpretation of experimental results. We generate 60 algorithmic designs to separate competing theories of choice under risk and show experimentally that designs with high separation generate stronger statistical results. We then analyze separation in a novel dataset of 146 lottery choice experiments from the literature, and provide evidence that algorithmically-constructed designs are better powered to separate popular theories of risk preferences. We further show that separation can be used to predict when an experimental design unfairly advantages one theory over its competitor. In an additional experiment, we compare statistical results across designs which appear similar, but structurally favor different theories. We observe predictable reversals in statistical conclusions, depending on how the design tasks are constructed.

*Thank you to Katie Coffman, Benjamin Enke, Anna Mikusheva, Sendhil Mullainathan, Matthew Rabin, Joshua Schwartzstein, and Jesse Shapiro for their invaluable advising and support. We also thank many Harvard Behavioral and Econometrics workshop participants for helpful comments and suggestions. We are grateful for financial support from the John and Kiendel Gordon fund. Jing: Department of Economics, Harvard Univeristy. bjing@hbs.edu. Shubatt: Department of Economics, Harvard University. cshubatt@g.harvard.edu.

1 Introduction

Empirical tests are a fundamental building block of theory. Designing insightful ways to test theory has been a core focus of experimental economics in particular. Typically, researchers rely on intuition or analytical insights to construct experimental tests. There are at least two problems with this. First, researchers may be leaving experimental power on the table if intuitive tests are inefficient from the perspective of statistical power. Second, intuitive tests may unfairly advantage some theories over others, which can lead researchers to misleading conclusions.

In this paper, we propose a design criterion for theory-testing experiments. Our criterion draws from the statistics and econometrics literature on experiment design, and it is tailored to the specific setting of experiments in behavioral economics. It captures the basic intuition that we value experimenting in spaces where competing theories make very different predictions. The greater the separation between theories on an experimental design, the better-powered the experiment is to reveal the true theory over its rivals. This criterion is straightforwardly useful at the experimental design stage, where researchers can use it to search systematically for highly-powered designs. The same theoretical structure allows us to characterize how choice of design may drive statistical results in non-obvious ways. This turns out to be useful at the analysis stage for identifying results which are likely to be power artifacts.

To fix ideas, consider a researcher wants to test Expected Utility Theory (EUT) by asking subjects to make two choices between lotteries. She considers the two designs in Figure 1.

Design X

Task 1

Option A	Option B
\$28 w.p. 100%	\$0 w.p. 5% \$28 w.p. 75% \$40 w.p. 20%

Task 2

Option A	Option B
\$0 w.p. 75% \$28 w.p. 25%	\$0 w.p. 80% \$40 w.p. 20%

Design X′

Task 1

Option A	Option B
\$30 w.p. 75% \$0 w.p. 5% \$40 w.p. 20%	\$0 w.p. 20% \$40 w.p. 80%

Task 2

Option A	Option B
\$30 w.p. 100%	\$0 w.p. 20% \$40 w.p. 80%

Figure 1: Choice tasks under Design X and Design X′.

On both designs, EUT predicts that subjects prefer the same option in both tasks. The Design X tasks are related by a common consequence, and the Design X' tasks are related by a common mixture. When we think about which design is a better test of EUT, we should ask: what behavior would we expect to see if preferences did *not* satisfy EUT? Concretely, we prefer designs where alternative theories of risk preference predict EUT-violating behavior. If instead alternative theories predict EUT-consistent behavior, then even if one of these theories is correct, we will never observe data which rejects EUT. Note that this problem can't be solved by simply collecting more observations – it is a fundamental limitation in what the design is capable of revealing.

Of course, there are many possible alternative theories, but a natural starting place is prospect theory (Kahneman and Tversky, 1979), which posits that individuals subjectively overweight small probability states and underweight large probability states when evaluating risky prospects. Prospect theory and it's successor cumulative prospect theory (Tversky and Kahneman, 1992) are some of the most widely-studied alternatives to EUT. If our goal is to test EUT, then we likely want a design which is powered against these alternatives.

It turns out that on Design X, prospect theory predicts behavior that violates EUT; and on Design X', prospect theory predicts behavior which is consistent with EUT. These predictions hold for a range of typically-estimated parameter values.[1] This means that before collecting any data on Design X', we know that even if decision-makers have prospect theory preferences, we will observe EUT-consistent behavior when we run the experiment. In this case, the experimental data tell us nothing about whether EUT describes the world well relative to Prospect Theory. If we instead run the experiment on Design X, then if Prospect Theory is right, the failure of EUT will be apparent.

This is not merely a theoretical issue. The Design X' tasks come from the experiment of Castillo et al. (2018), which is used to study risk preferences. Based on experimental data, the paper concludes "expected utility is as good as other models at explaining behavior."[2] Design X comes from Starmer and Sugden (1991), and in contrast, the experiment reveals that subjects are *not* consistent with EUT. These different conclusions need not be interpreted as fundamental inconsistencies in behavior; but instead as differences in what the experimental designs were set up to reveal.

Guided by this insight, we develop a formal criterion for experimental design which captures a design's ability to test between competing theories. This criterion is based on an

[1]The same is true for cumulative prospect theory.

[2]The set of other models considered includes cumulative prospect theory. The experiment includes several additional tasks which we have excluded here for expositional ease, but across the full set it remains true that behavior predicted by cumulative prospect theory is EUT-consistent for most tasks.

experiment's statistical power to correctly select the true theory over its rivals in a horse-race. A design is better according to this criterion if, supposing one theory is true, competing theories cannot explain the resulting data well. We then develop an algorithmic procedure for experimental design, wherein the researcher inputs a set of theories and a space of candidate designs and uses the criterion to optimize design. Because this optimization problem is generally non-convex and high-dimensional, we adopt new methods from the Bayesian Optimization literature to effectively search the space of designs.

We first demonstrate the usefulness of our criterion for constructing experimental designs using a combination of new experiments and a meta-analysis of the choice-under-risk literature. In our own experiments, we show that more highly-powered designs according to our criterion tend to produce stronger empirical results when we actually collect data. Furthermore, we assess the power of 146 designs from 95 different papers in the choice under risk literature and find that even the best designs appear to leave power on the table, under various assumptions on researchers' objectives.

In addition to guiding researchers at the experimental design stage, our criterion also offers new insight on ex-post analysis by highlighting how design features contribute to the results we see. For one, our criterion can identify when a design structurally favors one theory over another before any data is collected. In a second experimental exercise, we manipulate design parameters in seemingly subtle ways and show that (1) resulting designs often produce *opposite* conclusions on the same pair of theories when data is observed, and (2) our ex-ante criterion systematically predicts these varied conclusions. We then return to our collection of designs from the choice-under-risk literature, specifically focusing on papers which test Expected Utility Theory. We observe that many designs which fail to reject EUT conclusively are, according to our criterion, low-powered to test EUT against popular alternatives. Again, the results we observe do not strictly tell us about the relative value of theories; they also reflect limitations of design.

In Section 2, we define our design criterion and connect it to statistical power for model selection. Formally, we treat theories as model classes defined on some parameter space Θ. Fixing a particular value of θ, a model specifies a distribution of behavior for any potential experimental task. For instance, if the experimental domain is binary choice tasks between monetary lotteries, one relevant parametric model class might be CRRA expected utility with a tremble. Fixing a risk aversion parameter and a tremble probability, each model within this class specifies computable choice probabilities on every possible lottery pair.

We define a design's *normalized separation* between a fixed model (distribution of behavior) and a single rival class to be the normalized Kullback-Leibler (KL) divergence between

the fixed model and the best-approximating model (in terms of KL divergence) from the rival class. For multiple models, the definition is similar in spirit. If the fixed model were true, then the normalized separation along with the number of subjects would exactly determine the experiment's statistical power to select the correct model class over its rivals in a log likelihood horserace. Here, normalized separation plays the same role as the hypothesized treatment effect used in a standard power calculation. We think of normalized separation as capturing the component of an experiment's power which is attributable to the choice of design.

To calculate normalized separation of a fixed CPT model from EUT on a lottery choice design, we would (1) compute choice probabilities using the fixed CPT parameter values; (2) compute choice probabilities under the EUT parameter values that most closely match the CPT-choice probabilities in terms of KL-divergence; and (3) calculate the variance-normalized KL divergence between these two distributions, which we divide by the total number of tasks. In many cases, researchers will be interested in separability across a range of fixed models – this could be different parameter values under a single model class, or different model classes altogether. In such cases, we take the relevant measure to be a weighted average over the fixed models of interest. Our proposed design criterion ("design power") is a weighted average of normalized separation, with weights determined by a researcher-specified prior over plausible models. We value designs with a higher value of this criterion because we expect them to generate *stronger* statistical results when we use data to distinguish our models of interest.

We will sometimes be interested in separately analyzing separation and power depending on *which* model class we take as fixed. Instead of CPT and EUT, consider now two non-nested models of risk preference: CPT and salience theory of choice under risk (Bordalo et al., 2012). Both models assume that preferences are determined by subjective probabilities, but offer different accounts of how probabilities are distorted. It may be that a design is well-powered to separate CPT from salience, taking CPT as the fixed model; but not well-powered to separate salience from CPT. We call designs like this "asymmetrically powered," and we consider not only how statistical results vary with the average power across the two models, but the difference in power across the two models. We show that while average power has implications for the expected *strength* of statistical evidence, the difference has important implications for the *direction* of statistical evidence. Specifically, if one design is asymmetrically powered in favor of CPT, and another is asymmetrically powered in favor of salience, then we expect that results from the first design are more likely to indicate that CPT is the better model.

In Section 3, we discuss how to practically deploy our design criterion to generate new experimental designs. We treat design as an optimization problem, which requires the researcher to explicitly specify the space of potential designs over which to optimize. Restrictions on the design space may reflect feasibility constraints, such as payout maxima; or general researcher judgment, such as which regions of the domain are application-relevant. The researcher must also specify a prior over models. This prior could reflect the researcher's beliefs over the true data-generating process; but it can also be thought of as simply a practical way of encoding the models that the researcher cares to distinguish.

Having specified these objects, the problem of finding the "optimal" design according to the separation criterion remains computationally challenging for several reasons. First, our design criterion is non-convex, which means gradient-based methods will struggle to effectively traverse the landscape. Second, the design space is typically high-dimensional, which means exhaustive grid-based approaches are infeasible. Finally, the objective is costly to calculate, as each calculation requires an optimization sub-procedure to find the KL-minimizing parameters of the competing model classes. Making this calculation on a fixed design is easily manageable; but making the same calculation thousands of times throughout an optimization routine can be prohibitively slow.

In light of these challenges, we deploy Trust Region Bayesian Optimization (TuRBO) to algorithmically solve our design optimization (Eriksson et al., 2019). Bayesian optimization is a standard approach for handling costly objective functions, in which rather than compute the full objective with each optimization step, the routine probabilistically models surrogate objective which can be cheaply calculated. The surrogate is used to identify promising points in the search space, the full objective is computed on sufficiently promising points, the surrogate function updates in response to the new information, and the search continues. Because of the difficulty of our optimization problem, TuRBO cannot guarantee that we find the globally optimal point; but computational experiments show that it successfully discovers good designs more effectively than alternative optimization procedures.

In Section 4, we demonstrate the value of the criterion and algorithm for producing empirically useful designs, and we compare the resulting algorithmic designs with designs drawn from the existing literature. The primary purpose of the design criterion is to help us identify before data is collected which designs are likely to generate stronger empirical results. In Section 2, we show that normalized separation theoretically corresponds to statistical power under strong assumptions about the true data generating process; but in practice, our models may be misspecified, or our priors may be miscalibrated. In this case, we no longer have formal guarantees; so we want to assess empirically whether the criterion

is nonetheless a good ex-ante predictor of the strength of ex-post evidence.

We deploy our algorithmic design procedure to generate a set of 60 lottery choice designs for a range of objectives. We consider 6 distinct theories of choice under risk and generate algorithmic designs optimized for different design criteria corresponding to different theory-pairs. Some designs are optimized to separate a larger model (e.g. CPT) from a smaller model class (e.g. EUT); some are optimized to mutually separate two non-nested model classes; and some are optimized to distinguish one model class from its competitor, without being able to separate in the reverse direction.

We then run all 60 designs and empirically validate the separation criterion for different pairs of risk theories, corresponding to 8 different potential researcher objectives. For each theory-pair, we have three designs which are algorithmically generated to maximize the separation criterion between those two theories; and many other designs generated for other objectives, resulting in variation in the value of the criterion. We want to check that for each of these pairs, design power to separate the two theories positively predicts the strength of statistical results. For each design and for each theory-pair, we first the value of the separation criterion, and then use observed data to calculate the Vuong model selection z-statistic. We show that separation positively predicts the absolute value of this test statistic for all 8 theory-pairs, with 4 pairs significant at $p = 0.05$. To give a sense of magnitudes, on our 8-task designs with 30 subjects, moving from a design with median CPT-EU power to one with maximum power increased the (standard normal) z-statistic for rejecting EU in favor of CPT by 1.6 units.

After asserting that the separation criterion is meaningfully associated with the strength of empirical outcomes, we then assess the power of our algorithmic designs as compared to designs typical in the experimental literature. We construct a novel dataset of 146 distinct lottery choice experimental designs from 95 different papers and assess the ability of these designs to distinguish between popular competing models of risk. We manually classify papers according to which theories (if any) the experiments purport to test, and find that experiments testing EU and CPT (or another form of rank-dependent utility) on average have greater power to separate CPT from EU than experiments which do *not* make such a claim. This suggests that our criterion is broadly aligned with researchers' goals, and that researchers are able to recognize better designs over worse designs. However, our algorithmic designs far outperform even the *best* designs from the literature, both on CPT-EUT power and for the 7 other theory-pair criteria from our experiment. This apparent advantage is robust to variation in the parametric priors used to compute power. This indicates to us that there is meaningful scope for researchers to cheaply improve their experimental designs by

deploying our algorithmic procedure.

In Section 5, we demonstrate how our design criterion can also be useful for understanding where experimental results come from. One way to think about the criterion is that it's providing a language for articulating what an design is and is not set up to test. When an empirical exercise is not aligned with what a design is capable of testing, we can find ourselves with statistically valid but theoretically misleading results.

In a second set of experiments, we show how asymmetry in ex-ante theory separation generates spurious variation in statistical conclusions. A design is asymmetrically powered to separate two theories if, for instance, Theory 2 is unable to rationalize data generated by Theory 1; but Theory 1 can approximately fit data generated by Theory 2. We say such a design *asymmetrically favors* Theory 1; and if the reverse is true, we say the design asymmetrically favors Theory 2. In our experiment, we take six non-nested pairs of theories of risk preferences and generate two types of designs for each: designs which asymmetrically favor Theory 1, and designs which asymmetrically favor Theory 2. Then, as in our earlier experiment, we collect data on all the designs, fit both Theory 1 and Theory 2, and analyze their relative performance. Across all 6 pairs of theories, we find that empirical results are indeed more favorable to Theory 1 on the designs which asymmetrically favor Theory 1, as opposed to the designs which asymmetrically favor Theory 2. But furthermore, we often observe *reversals* in statistical conclusions as we move between the two types of designs. That is, we can have one design on which Theory 1 statistically significantly outperforms Theory 2; and another on which Theory 2 statistically significantly outperforms Theory 1. Importantly, these reversals are not random – they are predictable consequences of the shifting design structure.

Finally, we return to our collection of experimental designs from the lottery choice literature, narrowing to those which test Expected Utility Theory, and try to understand how considering design power leads us to new insights on how to interpret results. We document that the papers which came to a positive conclusions about EUT had relatively *poor* separation of CPT and EUT on their designs. This means that even if subjects exhibited CPT-like preferences, we wouldn't necessarily expect to see large deviations from EUT on these designs.

We also discuss the set of papers which document clear rejections of EUT, but do not have designs with high CPT-EUT separation. This indicates that although EUT fails on these designs, CPT is not a good description of how EUT fails. We analyze in particular the classical Allais paradox, which famously demonstrates a failure of EUT that is often attributed to CPT. We show that the Allais paradox is a low-power design according to our criterion.

This reflects that at standard parameter values, CPT does not predict the typical pattern of EUT violations, an incongruence first pointed out by Neilson and Stowe (2002). To be clear CPT *can* rationalize the paradox through subjective probability-weighting; but the degree of probability distortion needed to generate paradoxical behavior is not consistent with what we observe elsewhere. We think of this as an example of an *intuitive* test of EUT that can technically be rationalized under CPT; but it is not the test a researcher would design if they had CPT in mind.

Our framework for experimental design relates to several strands of prior literature. Formally, our criterion is closely related to prior work on optimal design in econometrics and statistics. This literature proposed choosing designs on which alternative models poorly fit data generated by the hypothesized true model, first for regression models, and later for arbitrary distributions using KL divergence to pseudo-true alternatives (Atkinson and Fedorov, 1975; López-Fidalgo et al., 2007). Our unnormalized separation measure corresponds to a prior-averaged version of this KL criterion, following Bayesian design strategies proposed by Ponce De Leon and Atkinson (1991); Tommasi and López-Fidalgo (2010); Dette et al. (2015).

Building on these established technical results, we aim to bridge the gap between the methodological literature and the practical needs of experimental researchers. First, we normalize the KL divergence to an interpretable standard-normal scale (closely following Vuong (1989)), which facilitates a clear interpretation with respect to statistical power for horseracing exercises standard in the literature. Second, we clarify a number of practical questions that arise in model selection among behavioral models, including the role of parameter identification, alternative model-selection practices, and extensions to heterogeneity analysis. Third, we port tools from the Bayesian Optimization literature to implement a practical procedure which researchers can use to generate new experimental designs. Finally, we derive empirically testable predictions linking the theoretical design-power criterion to the behavior of model-selection evidence observed ex post.

A number of papers in economics and cognitive science discuss objectives for experimental design and consider notions of efficiency and power specific to behavioral economics. Our design criterion is conceptually consistent with Bronars Power (Bronars, 1987) and restrictiveness (Fudenberg et al., 2023), which are both measures of how well alternative distributions of behavior accord with theory on a given design. We take our criterion one step further by turning it into an objective function for experimental design. Niederle (2025) and Snowberg and Yariv (2025) both posit that experimenters should look for designs where competing theories make maximally different predictions; our criterion formalizes this idea.

Chapman et al. (2024) and Bland (2025) provide approaches to optimal design for precisely estimating parameters under a fixed model, and Healy and Leo (2024) theoretically characterizes experiments which discriminate between deterministic theories. We instead focus on the distinct empirical objective of statistically discriminating between stochastic models. The most conceptually similar papers are Balietti et al. (2021) and Myung and Pitt (2009), both of which propose algorithmic design procedures for statistical model discrimination. Relative to these, we use a novel design criterion grounded in experimental power and provide an updated algorithmic approach suitable for substantially richer design spaces.

Finally, we relate to a literature on how algorithms can be deployed to discover theoretically interesting experiments. Both Fudenberg and Liang (2019) and Mullainathan and Rambachan (2023) propose using algorithms trained on existing behavioral data to generate new experiments on which known theories are predicted to fail. Our paper highlights how algorithmic search can improve design even in the absence of rich data on behavior. A researcher always needs *some* way of determining which data should be gathered, and we argue that the set of hypothesized models serves as a useful proxy for what the truth might look like. We show that this procedure works in the sense that the designs produced by our algorithmic procedure tend to produce stronger statistical evidence, which is exactly the experimenter's goal under our framework.

2 Theoretical Framework

We model a researcher choosing a set of experimental tasks to test between competing theories. The statistical power of an experiment depends on both its size and its component tasks. If two theories make nearly the same predictions on the experimental tasks, even a large dataset may not distinguish them. If the predictions are far apart, a small dataset can produce clear evidence. This section makes that idea precise, turns it into a practical design criterion, and derives empirical relationships which we test later on.

2.1 Model Separation and Statistical Power

We define an experimental design as a collection of T tasks. Let $\mathcal{X}$ be the task space and let $\mathbf{X} = (X_1, \ldots, X_T) \in \mathcal{X}^T$ denote a design. For example, a task may be a binary choice between two lotteries as in our example from Figure 1. In this case, an experiment is a collection of T lottery pairs. Let $\mathcal{Y}$ denote the behavioral response space associated with a task. For a generic binary choice task, $\mathcal{Y} = \{0, 1\}$. When an experiment is run, N subjects complete

each task in $\mathbf{X}$ and produce $n = NT$ observations, which we denote $\mathbf{Y} = \{Y_{it} : i \leq N, t \leq T\}$. For a binary choice experiment, we observe $Y_{it} \in \{0, 1\}$ for each subject i and each task t.

Formally, we treat theories as parametric model classes, and we model a researcher who will use experimental data to separate K competing classes. A model class M_k is a collection of models

$$M_k = \{g_k(y \mid x, \theta_k) : \theta_k \in \Theta_k\}, \qquad \Theta_k \subseteq \mathbb{R}^{\ell_k},$$

where a model (M_k, θ_k) specifies a distribution of behavior $Y \sim g_k(\cdot|X, \theta_k)$ for any task $X \in \mathcal{X}$. Each model assumes behavior is independent across individuals i and tasks t. The true model is fixed but is unknown to the researcher.

Example 1 (Risk Preferences in Lottery Choice)**.**
Consider a researcher who wants to test between Cumulative Prospect Theory (CPT) and Expected Utility Theory (EUT) using binary choices between lotteries. Her task space $\mathcal{X}$ is choices of the form:

Option A:	\$$w$ with probability p,	\$0 otherwise
Option B:	\$1 for sure	

She formulates the theories CPT and EUT as parametric model classes, where parameters pin down binary choice probabilities as a function of $X = (w, p)$. We demonstrate using Prelec probability weighting function for CPT (Prelec, 1998), and a power utility function and fixed tremble probability for both. Probability of choosing Option A is given by:

$$g_{CPT}(1 \mid X, \theta_{CPT}) = \begin{cases} 1-\epsilon, & \exp(-(-\ln p)^{\gamma}) \cdot w^{\alpha} \geq 1 \\ \epsilon, & \text{otherwise} \end{cases}$$

$$g_{EUT}(1 \mid X, \theta_{EUT}) = \begin{cases} 1-\epsilon, & p \cdot w^{\alpha} \geq 1 \\ \epsilon, & \text{otherwise} \end{cases}$$

where $\epsilon \in [0, \frac{1}{2})$ is the probability of a trembling-hand error and $Y = 1$ indicates whether the subject chooses Option A. ◊

In order to talk about an experiment's statistical power, we need to specify how experimental data will be analyzed. To fix ideas, we assume for now that the researcher uses likelihood-based model comparison. However, our ideas extend to generic statistical deci-

sion theory and other data analysis procedures, as discussed in Appendix A.1.[3]

Concretely, we assume that the researcher analyzes the relative performance of distinct parametric model classes by (1) fitting parameter values of each model to observed data using maximum likelihood estimation, and (2) comparing the likelihoods of the fitted models. This corresponds to horse-racing exercises common in the experimental literature (see e.g. Camerer and Ho (1994)). With two classes, M_1 and M_2, a positive log-likelihood ratio favors M_1 and a negative ratio favors M_2. A Vuong test uses the same log-likelihood ratio after dividing by its estimated sampling variation. This puts the comparison on a standard-normal scale, so its magnitude can be read against familiar critical values (Vuong, 1989).

We derive our design criterion from the expression for an experiment's power to correctly select model class M_1 over its competitors using this approach, in a world where the true model is (M_1, θ_1). To build intuition for the criterion, it's helpful to consider what we expect to happen in a likelihood horserace under this model. When we horserace competing parametric model classes, we do not compare likelihoods of the models at *arbitrary* parameter values; we compare likelihoods at the *estimated* parameter values. For model class M_1, the estimated parameter $\hat{\theta}_1$ converges to the true value θ_1. For any other model class M_k, the estimated parameter value $\hat{\theta}_k$ converges to θ_k^*, which is the value $\theta_k \in \Theta_k$ that allows M_k to most-closely approximate the predictions of (M_1, θ_1) on the dataset (White, 1982). [4] This object θ_k^* is called the *pseudo-true parameter* of M_k in the Econometrics literature, and it is a key input to our design criterion. We state the formal definition below.

Definition 1 (Pseudo-true model)**.**
Assuming the data on each design task X_t is generated by some distribution $p(\cdot \mid X_t)$, a pseudo-true parameter *is any*

$$\theta^*(p, X) \in \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}(p(\cdot \mid X_t) \,\|\, g(\cdot \mid X_t, \theta)),$$

where $\mathrm{KL}(f \| g) = \int f(y) \log[f(y)/g(y)]\, dy$*. The model* $(M, \theta^*(p, \mathbf{X}))$ *is the pseudo-true model of* p*.*

The pseudo-true model for p from model class M_j is the model that minimizes Kull-

[3]For example, experiments testing the Generalized Axiom of Revealed Preference (GARP) have traditionally used model selection rules counting strict GARP violations, which are not likelihood-based but can still be accommodated within generic decision theory.

[4]Convergence of the estimators depends on standard regularity assumptions required for maximum likelihood estimation.

back–Leibler divergence on the tasks in $\mathbf{X}$.[5] For $p = g_k(\cdot \mid \cdot, \theta_k)$, we abuse notation slightly by writing $\theta_j^*(\theta_k, \mathbf{X})$ to denote the pseudo-true parameter of M_j. The model $g_j(\cdot \mid X_t, \theta_j^*(\theta_k, \mathbf{X}))$ thus denotes M_j's best approximation of model (M_k, θ_k) on the tasks in design $\mathbf{X}$. Note that the pseudo-true model is design-dependent – even if the true model is held fixed, the pseudo-true model may change as we move between designs.

We next define the *separation* between a fixed model and a rival model class, which will be the basis of our design criterion.

Definition 2 (Separation and normalized separation)**.**
Fix a model (M_k, θ_k), a rival class M_j, and a design $\mathbf{X}$. For a single observation Y generated by the model on task $X_t \in \mathbf{X}$, define the log-likelihood ratio of (M_k, θ_k) and the pseudo-true model from M_j:

$$L_t^{kj} = \log \frac{g_k(Y \mid X_t, \theta_k)}{g_j(Y \mid X_t, \theta_j^*(\theta_k, \mathbf{X}))}, \qquad Y \sim g_k(\cdot \mid X_t, \theta_k). \tag{1}$$

The separation *of (M_k, θ_k) from M_j on $\mathbf{X}$ is*

$$\bar{D}_{kj}(\theta_k; \mathbf{X}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[L_t^{kj}].$$

The normalized separation *is*

$$\bar{\rho}_{kj}(\theta_k; \mathbf{X}) = \frac{\bar{D}_{kj}(\theta_k; \mathbf{X})}{\sqrt{\bar{V}_{kj}(\theta_k; \mathbf{X})}}, \qquad \bar{V}_{kj}(\theta_k; \mathbf{X}) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{Var}(L_t^{kj}). \tag{2}$$

For $K > 2$, let $\bar{\rho}_k \equiv (\bar{\rho}_{kj})_{j \neq k}$ and define $(\bar{R}_k)_{jj'} = \frac{1}{T} \sum_t \mathrm{Cov}(L_t^{kj}, L_t^{kj'}) / \sqrt{\bar{V}_{kj} \bar{V}_{kj'}}$.

$\mathbb{E}[L_t^{kj}]$ is the KL-divergence of the fixed model from the pseudo-true rival model on X_t, which makes *separation* $\bar{D}_{kj}$ the task-level average KL-divergence. When we compare theories by assessing the fit of parametric models, it's not the fit of an arbitrary model in M_j that matters – it's the fit of the *best* model in M_j. If the empirical distribution matched (M_k, θ_k), this would be $(M_j, \theta_j^*(\theta_k, \mathbf{X}))$. Separation is a measure of how much evidence an average observation from design $\mathbf{X}$ contributes to separating these two models.

The normalization in $\bar{\rho}_{kj}$ accounts for how variable that evidence is. Two designs can have the same average KL distance even though one produces steady evidence and the other relies on rare, large likelihood differences. Dividing by $\sqrt{\bar{V}_{kj}}$ expresses separation in units

[5] Note that the pseudo-true model is not in general unique, but for later results we will require all pseudo-true models to induce the same distribution of behavior on fixed tasks $X_t \in \mathbf{X}$.

of its own sampling noise and therefore distinguishes the two. When $K > 2$, $\bar{R}_k$ captures how the evidence against different rivals co-moves because all pairwise comparisons use the same outcomes.

Example 2 (Risk Preferences in Lottery Choice, Continued)**.**
Fix a CPT model with $\theta_{CPT} = (\gamma_0, \alpha_0, \tau_0)$. Define q_t to be the probability of choosing A on X_t under (M_{CPT}, θ_{CPT}), and q_t^* to be the probability of choosing A on X_t under the pseudo-true model from M_{EUT}:

$$q_t = g_{CPT}(A \mid X_t, \theta_{CPT})$$
$$q_t^* = g_{EUT}(A \mid X_t, \theta_{EUT}^*(\theta_{CPT}, \mathbf{X}))$$

The *separation* of (M_{CPT}, θ_{CPT}) from M_{EUT} on $\mathbf{X}$ is

$$\bar{D}(\theta_{CPT}; \mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} q_t \log\frac{q_t}{q_t^*} + (1-q_t)\log\frac{1-q_t}{1-q_t^*}. \tag{3}$$

The *normalized separation* is

$$\bar{\rho}(\theta_{CPT}; \mathbf{X}) = \frac{\bar{D}(\theta_{CPT}; \mathbf{X})}{\sqrt{\bar{V}(\theta_{CPT}; \mathbf{X})}}, \quad \bar{V}(\theta_{CPT}; \mathbf{X}) = \frac{1}{T}\sum_{t=1}^{T} q_t(1-q_t)\left[\log\frac{q_t(1-q_t^*)}{q_t^*(1-q_t)}\right]^2 \tag{4}$$

◊

Normalized separation is our design criterion of choice because of its formal connection to an experiment's statistical power for model selection. The following proposition makes this connection explicit. Given the data, let $\hat{L}_{kj}$ denote the empirical log likelihood ratio of the likelihood maximizing models from classes k and j. Let $\hat{k}$ denote the class with the largest fitted likelihood, and, fixing class k as the truth, define $\hat{\mathbf{D}}_k = \left(\hat{L}_{kj}\right)_{j\neq k}$ and $\bar{\mathbf{D}}_k = (\bar{D}_{kj})_{j\neq k}$ as vectors of empirical and theoretical separation. Let ϕ_R denote the CDF for a multivariate normal random variable $W \sim \mathcal{N}(0, R)$, i.e. $\phi_R(v) = \Pr(W \leq v)$, where the inequality $W \leq v$ is equivalent to $W_i \leq v_i$ for every component i.

Proposition 1 (Power of a design)**.**
Fix a data-generating model (M_k, θ_k) and a design $\mathbf{X}$ with T tasks. Suppose the following conditions hold:

(a) For every alternate model class M_j, every pseudo-true parameter θ_j^* induces the same distribution on the tasks in $\mathbf{X}$.

(b) For every $j \neq k$, $0 < \bar{V}_{kj} < \infty$.

(c) Every model class satisfies the likelihood regularity conditions stated in Appendix D.1, ensuring first-order equivalence between its maximized likelihood and the likelihood at its pseudo-true distribution.

Then, as $N \to \infty$,

$$\sqrt{NT}\left(\hat{\mathbf{D}}_k - \bar{\mathbf{D}}_k(\theta_k; \mathbf{X})\right) \xrightarrow{d} \mathcal{N}(0, \bar{\Sigma}_k),$$

where

$$(\bar{\Sigma}_k)_{jj'} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{Cov}(L_t^{kj}, L_t^{kj'}), \qquad (\bar{\Sigma}_k)_{jj} = \bar{V}_{kj}.$$

Consequently,

$$\Pr\big(\hat{k} = k \mid X, (k, \theta_k)\big) \approx \Phi_{\bar{R}_k}\big(\sqrt{NT}\,\bar{\rho}_k(\theta_k; \mathbf{X})\big). \tag{5}$$

For two classes, this is the ordinary univariate normal probability $\Phi(\sqrt{NT}\,\bar{\rho}_k)$.

Condition (a) allows the pseudo-true parameter to be non-unique, so long as all such parameters make the same predictions on the tasks in $\mathbf{X}$. Condition (b) requires each likelihood comparison to have finite, nonzero sampling variation, and in particular, no rival model can reproduce the generating distribution exactly on the design. These are standard regularity and separation conditions for likelihood comparisons, stated at the fixed data-generating model rather than imposed globally on the model classes.

Proposition 1 has a direct frequentist interpretation. If the true model were known to be (M_k, θ_k), then the experiment's statistical power to correctly select M_k over all $M_{j\neq k}$ as N goes to infinity is given by Equation 5. Power is increasing in N, the number of subjects; T, the number of tasks; and $\bar{\rho}_k$, the normalized separation. We think of normalized separation as the component of an experiment's power attributable to the choice of design. Although Proposition 1 is stated for maximum likelihood selection, the same first-order power expression applies when the researcher uses a penalized selection method such as AIC and BIC for fixed-dimensional models, provided that the penalty is on a scale of $o(\sqrt{NT})$ relative to the total log-likelihoods.

Importantly, for any fixed model and rival model classes, $\bar{\rho}(\theta_k; \mathbf{X})$ can be calculated *before seeing any data*. Under the hypothesis that data is generated under (M_k, θ_k), a design with higher normalized separation leads to greater experimental power.

This is the same logic as a textbook power calculation. For a one-sided test based on $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2/n)$,

$$\Pr(\text{reject } H_0 \mid \beta) \approx \Phi\left(\sqrt{n}\,\frac{\beta}{\sigma} - z_\alpha\right). \tag{6}$$

The normalized treatment effect β/σ determines how quickly power rises with sample size. Here $\bar{\rho}_k$ plays the same role: statistically decisive conclusions are easier when the design makes the relevant normalized separation large.

2.2 Design Criterion and Model Uncertainty

If the researcher knew the true model, she would simply choose the design tasks that maximize normalized separation from rival classes. The practical difficulty is that the true model is not known—this is why the experiment is needed. This is true in standard power calculations as well: in order to calculate power, a researcher must specify the effect size they want to be powered for, often called a "minimal detectable effect." In our case, the researcher will specify the models they want to be able to separate from one another. To simplify notation, the remainder of this section considers the case of selecting between two models ($K = 2$), denoted M_1 and M_2.[6]

Our proposed design criterion requires the researcher to supply a hierarchical prior over parameters within a model class, and over the model classes themselves. First, for both model classes M_k, the researcher supplies a prior over parameter values $\pi_k(\theta_k)$. We denote the prior-averaged normalized separation in each direction by

$$\bar{\rho}_1(\mathbf{X}) = \int_{\Theta_1} \bar{\rho}_{12}(\theta;\mathbf{X}) \cdot \pi_1(\theta)\, d\theta, \qquad \bar{\rho}_2(\mathbf{X}) = \int_{\Theta_2} \bar{\rho}_{21}(\theta;\mathbf{X}) \cdot \pi_2(\theta)\, d\theta.$$

Second, the researcher specifies prior weights over the model classes $P_1, P_2 \geq 0$, with $P_1 + P_2 = 1$. These are weights chosen in advance by the researcher, not estimates formed from the new data. The design objective is

$$\mathbf{X}^* \in \arg\max_{\mathbf{X}\in\mathcal{X}^T} \bar{\rho}(\mathbf{X}), \qquad \bar{\rho}(\mathbf{X}) = P_1 \cdot \bar{\rho}_1(\mathbf{X}) + P_2 \cdot \bar{\rho}_2(\mathbf{X}). \tag{7}$$

If we interpret P_1, P_2 and π_1, π_2 as encoding the researcher's Bayesian belief over the true model, then the design criterion $\bar{\rho}(X)$ is the expected normalized separation of the true model from the alternative, where the expectation is taken over hierarchical beliefs over the true model and its parameters.

Our criterion is closely related to prior work on optimal design for model selection. For a fixed scenario (M_k, θ_k), $\bar{D}_{kj}(\theta_k;\mathbf{X})$ is the KL discrimination criterion studied by López-Fidalgo et al. (2007) and averaging $\bar{D}_{kj}$ over generating parameters yields the Bayesian cri-

[6]For general K, see Appendix C.

teria studied by Ponce De Leon and Atkinson (1991); Tommasi and López-Fidalgo (2010), and Dette et al. (2015). Our objective $\bar{\rho}(\mathbf{X})$ instead averages over the normalized likelihood-ratio following Vuong (1989), and therefore puts KL separation on the standard-normal scale that, under the conditions of Proposition 1, governs model-selection power.[7]

Example 3 (Restrictiveness on Initial Play in Matrix Games)**.**
Consider a researcher studying initial play in 3x3 matrix games. She wants to assess the Logit Level-1 model from Fudenberg et al. (2023) et al. in an experiment, and she wants her set of experimental games to restrict the model's flexibility as much as possible. Let M_1 be the model class of *eligible* behaviors, defined as all distributions over the 3 actions such that strictly dominant moves are played with at least $1/3$ probability and strictly dominated moves played with at most $1/3$ probability. Let M_2 be the Logit Level-1 model. In this case, M_1 encodes an agnostic prior over models which the researcher would like to test against. We propose setting $P_1 = 1$, $P_2 = 0$, and placing equal weight on each distribution in M_1. We use the resulting average normalized separation of M_1 from rival class M_2 as the design criterion.

Note that in this setting, separation in our framework is closely analogous to the restrictiveness criterion proposed by Fudenberg et al. (2023). Our preferred criterion is prior-averaged normalized separation $\bar{\rho}$ due to its connection to standard inference critical values, but we observe that the average unnormalized separation $\bar{D}$ coincides with the numerator of the restrictiveness criterion proposed by Fudenberg et al. (2023) up to orientation of the KL divergence. In practice, the two criteria often induce a similar order over designs. The primary difference is how we think about what these objects represent: Fudenberg et al. (2023) emphasizes restrictiveness as a property of theory, while we emphasize (normalized) separation as a property of a design. ◊

2.3 Empirical Evaluation

The design criterion uses stylized models and hypothesized parameters. Its practical value depends on whether ex ante separation predicts the evidence that real data produce. Here we derive that link under the maintained model, and later we will test empirically whether it survives approximation and misspecification.

[7]There is a well known degeneracy issue with the normalized likelihood ratio when Condition (b) of Proposition 1 fails, we argue that the degenerate case is not particularly relevant for the purpose of power calculation. Section 2.4 discusses this in more details.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the maximum-likelihood estimates for the two models and define

$$\hat{D} = \frac{1}{NT} \sum_{i,t} \log \frac{g_1(Y_{it} \mid X_t, \hat{\theta}_1)}{g_2(Y_{it} \mid X_t, \hat{\theta}_2)}, \qquad \hat{z} = \frac{\sqrt{NT}\,\hat{D}}{\hat{\sigma}}. \tag{8}$$

Here $\hat{\sigma}^2$ is the sample variance of log-likelihood ratios. $\hat{D}$ is the mean log-likelihood ratio, and $\hat{z}$ is its Vuong standardization. Both are common model-selection statistics in empirical work. For the mean log-likelihood ratio, define the weighted directional quantities

$$\mathbb{D}_1(\mathbf{X}) = \int_{\Theta_1} \bar{D}_{12}(\theta; \mathbf{X}) \cdot \pi_1(\theta)\, d\theta \qquad \mathbb{D}_2(\mathbf{X}) = \int_{\Theta_2} \bar{D}_{21}(\theta; \mathbf{X}) \cdot \pi_2(\theta)\, d\theta.$$

Corollary 1 (Ex ante separation predicts ex post evidence)**.**
Under the conditions of Proposition 1:

(i) Conditional on truth $(1, \theta_1)$,

$$\hat{D} = \bar{D}_{12}(\theta_1) + O_p(n^{-1/2}), \qquad \hat{z} = \sqrt{n}\,\bar{\rho}_{12}(\theta_1) + O_p(1).$$

Under truth $(2, \theta_2)$, the centers are instead $-\bar{D}_{21}(\theta_2)$ and $-\sqrt{n}\,\bar{\rho}_{21}(\theta_2)$.

(ii) If the model and scenario weights form a hierarchical prior and the supported scenarios have enough moments for expectations to converge, then

$$\mathbb{E}_w[\hat{D} \mid X] \longrightarrow P_1\mathbb{D}_1 - P_2\mathbb{D}_2, \qquad \mathbb{E}_w[|\hat{D}| \mid X] \longrightarrow P_1\mathbb{D}_1 + P_2\mathbb{D}_2,$$
$$\frac{\mathbb{E}_w[\hat{z} \mid X]}{\sqrt{n}} \longrightarrow P_1\bar{\rho}_1 - P_2\bar{\rho}_2, \qquad \frac{\mathbb{E}_w[|\hat{z}| \mid X]}{\sqrt{n}} \longrightarrow P_1\bar{\rho}_1 + P_2\bar{\rho}_2.$$

Part (i) states that, for a fixed truth, each ex post statistic is centered on its matching ex ante quantity design by design. Part (ii) gives the corresponding result after averaging over scenarios. These statements suggest that, when the true model is known or the scenarios of models and parameters are exactly correctly specified, we should obtain a slope of one when we regress the ex post statistic on the ex ante power measure. Of course, we do not expect these conditions to hold in practice, hence our empirical benchmark is therefore a *positive* slope, not a slope of one.

More importantly, whether the slope is positive answers the question of whether by choosing powerful designs ex ante, we predictably produce stronger statistical evidence. This is the right test for evaluating whether our ex ante power measure is informative for choosing designs, rather than a test of whether the parametric models, parameter grid, or

scenario weights correctly describe the population.

Both sign and magnitude matter. The sign of $\hat{D}$ identifies the better-fitting model; $|\hat{D}|$ measures the strength of its advantage. The Vuong statistic has the same sign and expresses magnitude in estimated sampling-noise units. Thus directional separation predicts which model wins, while symmetric separation predicts how decisive the comparison is.

2.4 Discussion

Parameter ranges in overlapping and nested model classes. In general, we allow model classes to overlap or be nested. However, we argue that the parameters that conceptually make sense for power calculations lie outside the overlapping region, since within the overlapping region the two model classes induce identical distributions of behavior and are indistinguishable on any design.[8] As a result, we emphasize that that Proposition 1 is pointwise in the generating model and parameter. Condition (b), which requires $\bar{V}_{kj} > 0$, exactly corresponds to restricting our attention to parameters outside the overlapping region, where normalized separation and its power interpretation are well defined.

A separate subtlety regarding overlapping models lies in the downstream empirical analysis. Since maximum-likelihood model selection and Vuong tests are illustrative of common model-selection exercises, we use them as leading motivation for our power criterion in this section. However, both procedures require care in interpretation when model classes are nested or overlapping.[9] Since our preferred objective is defined at the normalized-separation $\bar{\rho}$ level, these subtleties in downstream inference are not inherent to our framework. Researchers could instead use alternatives such as penalized likelihood selection or modified model-selection statistics (Shi, 2015; Schennach and Wilhelm, 2017), and our ex ante power measure remains well motivated for these alternative downstream inference procedures.

Normalized separation as design criterion. In Section 2.2, we recommended choosing a design based on prior-weighted normalized separation $\bar{\rho}(\mathbf{X})$. One could instead maximize

[8] For example, in the case of EU versus CPT, a researcher trying to design a well-powered experiment necessarily cares about probability-weighting parameter γ away from 1, and the relevant notion of separation is the one-directional $\bar{\rho}_{\text{CPT}}$, as discussed in Example 1.

[9] It is well known that unpenalized maximum likelihood mechanically weakly favors the containing class and that the usual normal approximation for the classical Vuong statistic is not uniformly valid near the overlap of model classes.

prior-weighted average power at a chosen sample size,

$$\text{WAP}(X;n) = P_1 \int_{\Theta_1} \Phi\big(\sqrt{n}\,\bar{\rho}_{12}(\theta;\mathbf{X})\big)\cdot \pi_1(\theta)\,d\theta + P_2 \int_{\Theta_2} \Phi\big(\sqrt{n}\,\bar{\rho}_{21}(\theta;\mathbf{X})\big)\cdot \pi_2(\theta)\,d\theta. \quad (9)$$

We recommend (7) because it does not require N, and hence $n = NT$, to be settled early in the design process. Furthermore, note that if we worked with (9), we would place little value on additional separation when experimental power is already high, since the normal CDF flattens as its value approaches 1. In practice, additional separation is likely still useful. The following result states when the two criteria give the same ranking.

Lemma 1 (Ranking designs away from the power ceiling)**.**
Let $\bar{\rho}(X_A) > \bar{\rho}(X_B) > 0$, fix $\beta \in (1/2, 1)$, and let $z_\beta = \Phi^{-1}(\beta)$. If

$$1 - \frac{\bar{\rho}(X_B)}{\bar{\rho}(X_A)} \geq \frac{z_\beta^2}{6},$$

then $\text{WAP}(X_A;n) > \text{WAP}(X_B;n)$ at every n for which every scenario-specific power under X_A is below β.

Average normalized separation and average power need not rank every pair of designs identically, however, Lemma 1 isolates the comparisons that matter most for improvement: when the designs' power has not saturated and the gain in effect size is large, the rankings agree. Once some scenarios are already near certain selection, the ranking can depend on the chosen n because extra separation barely changes their power. In the following sections, we will abuse terminology slightly and refer to designs with high average normalized separation as highly powered designs for expositional simplicity.

Heterogeneity. Throughout this section, we discussed model selection under the assumption that subjects are homogeneous in their behaviors. This, of course, does not hold true. Researchers are often interested in analyzing heterogeneity across subjects and thinking of model selection more broadly as an exercise in studying the population shares of agents that are best described by different models. Appendix B provides a set of formal results and detailed discussion about common statistical procedures used to analyze heterogeneity and how our notion of design power extends to those settings.

The main takeaway is that design power typically matters for the statistical properties such as bias and variance of population share estimators. Importantly, using low-powered

designs may induce asymptotic bias in these estimators that cannot be eliminated simply by increasing the number of subjects.

Model misspecification. For the purpose of choosing well-powered designs, Proposition 1 provides a literal power interpretation under the assumption that one of the specified models generates the data. This result provides an analytical benchmark, but in practice it is likely that *all* of the researcher's models are misspecified. In this case, suppose $p_{0t} = p_0(\cdot \mid X_t)$ denotes the actual distribution of behavior on task X_t, and let $g^*_{jt} = g_j(\cdot \mid X_t, \theta^*_j(p_0, \mathbf{X}))$ denote class j's pseudo-true distribution on that task. Under analogous likelihood regularity conditions, we have $\hat{D} \xrightarrow{p} \Delta_0(\mathbf{X}) \equiv \frac{1}{T}\sum_{t=1}^{T}\left\{\mathrm{KL}(p_{0t}\|g^*_{2t}) - \mathrm{KL}(p_{0t}\|g^*_{1t})\right\}$, so likelihood selection favors the class that is closer in average KL divergence to actual behavior on the tasks in $\mathbf{X}$ (Vuong, 1989). Importantly, note that the task-level true distributions p_{0t} and the KL gaps can change with $\mathbf{X}$. Under arbitrary misspecification, the best-approximating model class among the parametric models can therefore change across designs, even for a fixed population of subjects. Consequently, conclusions about whether one model class is better than another should be defined relative to the specific design that produced the empirical evidence, and we caution against interpreting model-comparison results as "global" rankings that extend beyond that design.

3 Implementation

In order to generate designs guided by normalized separation as in Equation Equation (7), we need to set a prior over models and specify the design space $\mathcal{X}^T$. Then, we need a way of solving the resulting numerical optimization. In this section, we offer guidance on how researchers should specify the theories, priors and design parameters; discuss practical challenges that arise in optimization; and describe our preferred optimization approach, which utilizes Trust-region Bayesian Optimization (Eriksson et al., 2019). We also present an example of a lottery choice design produced by our search algorithm.

3.1 Specifying Model Priors and Design Space

The optimal design according to our criterion depends on (1) the domain of potential designs and (2) the priors over models, both of which are set by the researcher. These are of course large degrees of freedom, but articulating these explicitly gives the researcher a transparent way of communicating the assumptions justifying the design.

The design space $\mathcal{X}^T$ will typically incorporate three types of constraints. First, there are feasibility constraints – some tasks may simply not be feasible to run in the lab. In lottery choice, for example, a researcher may not have the budget to incentivize choices with large payouts. Second, there are subject comprehension constraints. If tasks are too complicated, subjects may not be able to provide reliable data. This may mean a researcher wants to avoid lotteries with many different states or unusual probabilities. Finally, there are theory-domain constraints. The researcher may simply not be interested in testing theories in some regions of the design space if those regions are not relevant for applications. For instance, they may wish to avoid lotteries with extremely small probabilities, where theories like cumulative prospect theory are known to make odd predictions.

The hierarchical prior over model classes and parameters may encode the researcher's beliefs about the true data generating process. These can be guided, for instance, by estimates from prior literature, or economic calibration assumptions. In many cases, the prior may instead simply be a practical way of encoding for which models the researcher would like to be powered. For instance, some researchers may only care about power to reject M_2 if M_1 is true, in which case they would set $(P_1, P_2) = (1, 0)$. Others may only care about separating the models in certain regions of the parameter space.

For the purposes of generating designs, we recommend using discrete parameter grids in model priors. This means that for each model class M_k, the prior can be described by a finite set of parameters $(\theta_{ks})_{s=1}^{S}$ and associated nonnegative weights (w_{ks}) that sum to 1. We can calculate $\bar{\rho}(\mathbf{X})$ for any type of prior, but a discrete grid allows for faster computation which is valuable for optimization. In our experiments, we use fairly coarse prior grids and find that the ex-ante design criterion tracks ex-post empirical results well. Below, we return to our running example to demonstrate how we might specify priors for risk preference models.

Example 4 (Risk Preferences in Lottery Choice, Continued)**.**
For nested classes like CPT and EUT, it is natural for the researcher to set model class priors $P_{CPT} = 1$ and $P_{EUT} = 0$. The large class can mechanically fit any data generated by a model in the small class; so the scenario in which we hope to separate the models is when the truth is outside the small class. Intuitively, we want a design where EUT is unable to fit the data well if there is truly probability distortion as in CPT.

The researcher then sets a prior over CPT parameters based on which models within the CPT class she wants to be able to separate from EUT. The probability weighting parameter γ in CPT is typically estimated between 0.4 and 1, where $\gamma = 1$ gets back to EUT. The power utility parameter α is typically estimated between 0.6 and 1, with 1 being risk neu-

tral. Suppose she wants to be able to separate CPT from EUT under moderate or extreme probability weighting, with moderate or no risk aversion and moderate trembling. Then, she could set π_{CPT} using the prior grid in Table 1 below and applying equal weights to each row. She does not need to specify a parameter prior π_{EUT} for EUT, since P_{EUT} is equal to 0.

γ	α	ϵ
0.6	1	0.25
0.6	0.7	0.25
0.8	1	0.25
0.8	0.7	0.25

Table 1: Grid of θ_{CPT} values to be used in design objective.

For this researcher, the design criterion $\bar{\rho}(\mathbf{X}) = \bar{\rho}_{CPT}(\mathbf{X})$ is the average normalized separation defined in Equation 4, averaged over the four θ_{CPT} values in Table 1. ◊

3.2 Design Search with Trust-Region Bayesian Optimization

Several features of our design problem pose challenges for standard optimization methods. First, the objective typically has a complicated global geometry. Fixing a design and a candidate model, the value of the objective is given by the normalized separation between the candidate model and its best-fitting alternatives, and since the parameters and possibly the class of the best fitting alternative change with the design, the resulting objective value as a function of design can be non-concave, strongly anisotropic, and non-smooth. Hence standard gradient-based procedures are often sensitive to initialization values and converge to low quality local solutions.

Second, computing the objective is itself expensive. Given a design, we need to first calculate behavioral predictions under all maintained models; and then we need to run an optimization sub-routine to find the best approximating model(s) from the rival class(es). Computing this once for a fixed design is usually easy, but hundreds of iterative computations may be infeasible.

Finally, the design space is often high-dimensional. Consider a simple binary choice task between two lotteries, each with two states. This single task may have as many as six dimensions: two payoffs and one probability for each lottery. Even a 10-task design like this would have sixty dimensions – and this would be relatively small by experimental standards. The large size of the search space makes grid-based approaches infeasible.

Given these challenges, we adopt a Bayesian-optimization approach tailored to high-dimensional, multimodal objectives: trust-region Bayesian optimization (TuRBO, Eriksson

et al., 2019.) Below, we describe the procedure at a high level. See Appendix E for precise mathematical details.

Bayesian optimization (BO) is a standard tool for optimization problems with costly objectives. Rather than optimize directly on the costly objective, BO works by probabilistically modeling the true objective. This probabilistic "surrogate model" is initialized with some prior over the objective at every point in the search space. An acquisition rule is deployed to select a promising design $\mathbf{X}$ for evaluation. The acquisition rule favors designs which, under the surrogate model, have either high mean, i.e. are expected to be good on average; or high variance, which may produce high draws. The costly objective is evaluated on $\mathbf{X}$, the surrogate model is updated in a Bayesian way using the learned value, and the process repeats.

A known difficulty of conventional global BO in high-dimensional, multimodal problems is that a single surrogate may fail to approximate the entire space well, inducing large posterior uncertainty that push towards over-exploration rather than actually improving the objective value. Inspired by TuRBO (Eriksson et al., 2019), we restrict candidate search to adaptive local trust regions. A "trust region" is a hyperrectangle in the design space centered at the best design found within it, and our procedure maintains one or more trust regions in the search process. At each iteration, we fit a Gaussian-process surrogate to the evaluations made in the local round and selects the next evaluation points by Thomspon sampling from the Gaussian-process posterior over the trust region. If several consecutive evaluations improve upon the region's best evaluation, the trust region expands; if several consecutive proposals fail to improve, it contracts. When a region shrinks below a minimum side length, it is terminated and reinitialized elsewhere (e.g., seeded near a new promising point), which helps avoid spending the full budget refining an unpromising basin.

Our implementation of design search deploys TuRBO with a single trust region. For full details on our surrogate model and acquisition rule, see Appendix E. In computational experiments, this produced better results more quickly than alternatives methods (see Appendix Figure 9).

3.3 Discussion

While TuRBO is our preferred approach, we make several caveats. First, we note that because of the non-convexity and generality of our optimization problem, we cannot guarantee global optimality of the returned solution. We think such a guarantee is not essential to our goal of using algorithmic search to produce designs with high power. As long as the

search substantially improves power relative to designs obtained otherwise, we view it as a helpful tool.

Second, a researcher's particular objective may render other optimization approaches more useful. For instance, Mullainathan and Rambachan (2023) pose an optimization problem for generating experimental "anomalies" which has the same structure as ours, and thus many of the same practical issues. However, they are specifically interested in *minimal designs*, e.g. simple pairs of choice problems which illustrate a theory failure. In their implementation on pairs of lottery choices, they further reduce the dimension of the search space by fixing a subset of features and successfully optimize using a combination of grid-based and gradient-based approaches. We recommend TuRBO because it is robust to many common issues that we anticipate; but practitioners may find other optimization algorithms are sufficient for their specific environment.

3.4 Example Design

For the remainder of the paper, we turn to the domain of choice under risk to deploy our design framework. In Figure 2, we present a simple, two-task design optimized to separate Cumulative Prospect Theory from Expected Utility Theory.[10] The task space $\mathcal{X}$ is binary choices between two-outcome lotteries in a common state space.

Choice 1

Probability	Option A	Option B
0.85	\$99	\$74
0.15	\$0	\$99

Choice 2

Probability	Option A	Option B
0.09	\$99	\$0
0.91	\$0	\$4

Figure 2: Two-task algorithmic design, highly-powered to distinguish CPT from EUT

To see why this design is highly-powered for this purpose, first look at Choice 1. Option A is a high chance of \$99, and a low chance of 0. Option B is a high chance of \$75, and a low chance of \$99. Option A has a higher expected value, but under CPT, subjects will overweight the state which pays \$0 with 15% and prefer Option B. EUT can accommodate this choice, but it requires the decision-maker to be moderately risk averse.

[10]The design objective uses logistic noise specifications of both models and is described explicitly in Section 4.

In Choice 2, Option A is a low chance of $99 with $0 otherwise, and Option B is a high chance of $4 with $0 otherwise. Now, CPT predicts that the high chance of $0 in Option A is deflated, and the low chance of $0 in Option B is inflated. This means that CPT predicts Option A is preferred to Option B. Notice that in the previous choice, EUT could accommodate the CPT prediction by increasing risk aversion; but in this choice, the CPT prediction is the riskier option. EUT can make sense of either choice individually; but can't make sense of the joint preference pattern predicted by CPT.

4 Separation and Design

To demonstrate the importance of design power when selecting experimental tasks, we turn to perhaps the most widely studied experimental paradigm in behavioral economics: binary choice between lotteries. We first want to assess the value of theory separation in leading us to empirically useful designs. To do this, we algorithmically generate and then run 60 new lottery choice experimental designs which vary in average separation for a range of risk preference theories. Then, we want to examine how our algorithmic designs compare to what researchers typically run. We collect a dataset of 146 unique experimental designs from 95 papers and evaluate them in terms of theory separation across a range of potential design objectives.

4.1 Algorithmically-Generated Designs

We use our algorithm to construct 60 lottery choice designs optimized for 20 distinct objectives (3 unique designs for each objective). A "design" in our experiment is a set of 8 choice tasks between two money lotteries. We decided on 8-task designs because we wanted designs large enough to identify model parameters; but small enough for an analyst to interpret which features of designs cause models to fail. In each task, both lotteries are defined in a common state space with three possible states. State probabilities are restricted to multiples of 0.05, and both lotteries' payoffs are restricted to positive whole dollars between 0 and 50. The common state structure is explained to subjects and communicated visually in the experimental implementation. When we construct algorithmic designs, we search over sets of 8 binary choice tasks which meet these criteria.[11]

[11]We also add a requirement that no more than 2 tasks involve first-order stochastic dominance, which gives us more reliable identification in numerical checks.

4.1.1 Design Objectives

Each of our 20 design objectives involves 2 competing theories of stochastic choice under risk. We consider a set of 6 models in total, described in brief below. Each model, when parameterized, specifies a binary choice probability for any lottery task. Full parametric specifications for these models can be found in Appendix F.1.

1. **Expected Utility with Logit Noise (EU).** We assume a power utility function, i.e. $u(x) = x^{\alpha}$.

2. **Cumulative Prospect Theory with Logit Noise (CPT).** We again assume power utility, and we add a Prelec 1-parameter probability weighting function.

3. **Random Expected Utility with a Tremble (REU).** We assume a population of power utility parameters α which follow a lognormal distribution, and a fixed tremble probability.

4. **Random Cumulative Prospect Theory with a Tremble (CPT).** We again assume a lognormal distribution of power utility parameters with a fixed tremble probability, and we add a (non-random) Prelec 1-parameter probability weighting function.

5. **Comparison Complexity (CC).** We use the expected utility CDF-complexity model from Shubatt and Yang (2024), again assuming power utility.

6. **Salience with Logit Noise (SAL).** We use the continuous salience value function proposed by Bordalo et al. (2012).

Each objective used to generate our designs is associated with a pair of theories (model classes) from this list. For two model classes M_1 and M_2, recall that $\bar{\rho}_1$ denotes prior-averaged normalized separation of M_1 from M_2, which approximately corresponds to statistical power to select M_1 over M_2 if M_1 is the truth; $\bar{\rho}_2$ denotes the reverse. Both of these objects depend on which parameter values we hypothesize as the "truth," which we discuss more below.

For most applications, we anticipate researchers will be interested in using one of two objective types, which we call *one-sided* power and *symmetric* power. We use *one-sided* power when M_1 and M_2 have a nesting relationship, such as with CPT and EU. Intuitively, if behavior is truly EU, CPT will be able to fit the data well on *every* design. From a theory-testing standpoint, the right notion of design power here is one-sided: we just want $\bar{\rho}_{CPT}$ to be as large as possible on our design, meaning EU will do a poor job trying to fit data from

CPT agents. Concretely, for theory pairs where M_1 nests M_2, we used $\bar{\rho}_1$ as the design objective. For M_1, M_2 without a nesting relationship, we usually care about *symmetric* design power, meaning we want both $\bar{\rho}_1$ and $\bar{\rho}_2$ to be as large as possible. We implemented this by maximizing $\min\{\bar{\rho}_1, \bar{\rho}_2\}$.[12]

For a second experimental analysis which we discuss more in Section 5, we were also interested in studying designs with *asymmetric* power on non-nested pairs. Specifically, we wanted to see how results varied across designs with large $\bar{\rho}_1$ and small $\bar{\rho}_2$, versus designs with large $\bar{\rho}_2$ and small $\bar{\rho}_1$. We generated designs with this feature this by maximizing either $\bar{\rho}_1 - \bar{\rho}_2$, or $\bar{\rho}_2 - \bar{\rho}_1$.

The theory pairs and associated design objectives are laid out in Table 2 below. By convention, we always label the alphabetically-first theory as Theory 1. In total, we have 20 distinct objectives, and we generate 3 different designs using each.

Theory 1	**Theory 2**	**One-Sided** $\bar{\rho}_1$	**Symmetric** $\min\{\bar{\rho}_1, \bar{\rho}_2\}$	**Asymmetric 1** $\bar{\rho}_1 - \bar{\rho}_2$	**Asymmetric 2** $\bar{\rho}_2 - \bar{\rho}_1$
CPT	EU	✓	-	-	-
RCPT	REU	✓	-	-	-
CC	CPT	-	✓	✓	✓
CC	RCPT	-	✓	✓	✓
CC	SAL	-	✓	✓	✓
CPT	RCPT	-	✓	✓	✓
CPT	SAL	-	✓	✓	✓
RCPT	SAL	-	✓	✓	✓

Table 2: Power objectives maximized in construction of lottery choice designs

4.1.2 Hypothesized Parameters

To compute the value of $\bar{\rho}_1$ and $\bar{\rho}_2$ for any given 8-task design, we need to specify priors over parameter values for M_1 and M_2. This is, in one sense, a large degree of freedom; a researcher may hypothesize any parameter values they like. But if the researcher is completely miscalibrated, the proof will be in the pudding: highly powered designs will *not* do a good job of distinguishing between competing theories. At the same time, hypothesized parameters need not be exactly right for these constructs to be useful. Recall that the researcher's goal is to find designs which reveal clear winners; so if highly-powered designs do this, it's an indication that the calibration is good enough to be valuable.

[12]Another researcher may reasonably prefer an alternative approach to combining $\bar{\rho}_1$ and $\bar{\rho}_2$, such as a sum. We found that using a minimum produced more balanced designs, which we prefer for our purposes.

For each model, we set discrete parameter priors guided by the previous literature. Some parameters have been estimated in many previous papers, such as the CRRA power utility parameter α, or the Prelec probability-weighting parameter γ. For these, we chose some benchmark values from the literature. For others, such as the continuous salience parameter δ from Bordalo et al. (2012), we could not find existing parametric estimates. Here we conducted calibration exercises using examples from the original paper and selected values consistent with these examples. To calibrate logit noise parameters, we fixed different preference parameters and used existing lottery choice datasets to come up with an appropriate level of noise. Since we hypothesize multiple parameter combinations for each model, we calculate $\bar{\rho}_i$ by taking a simple average over design power for each row in the parameter grid.[13] The grid for each model was preregistered[14] and can be found in Appendix F.1.

4.1.3 Experimental Procedures

We implemented our lottery choice tasks in a Qualtrics survey and recruited subjects through Prolific. We collected choice data on 480 distinct tasks (8 tasks from 60 different designs). After reading instructions and passing comprehension checks, our subjects encountered a random subset of 6 designs, or 48 tasks in total. They were not aware of the "design" groupings, but were simply told that they would encounter 48 choice tasks. Of the 521 subjects who passed our comprehension checks, 308 completed the entire survey and passed two attention checks to be included in our final sample. The number of subjects per task ranged from 26 to 38, with a median value of 31.

We explained the tasks to subjects by telling them that they had a choice between two lotteries, A and B, which both pay out real money depending on the color of a randomly drawn ball (red, blue, or green). We told them there would always be 20 balls in total, and that each task screen would inform them of the ball colors for that task and the payout amounts for both options. See Figure 3 for an example of a task display. Subjects were paid a $4.50 completion fee and informed that they would have a one-in-ten chance at being selected for a bonus payment if they successfully passed all attention checks. If selected for a bonus, one of their 48 choices would be selected at random, their chosen lottery would be played out by the computer, and they would receive the monetary payout. Full instructions can be found in Appendix F.2. The median bonus payment was $39 and the median completion time was 22 minutes.

[13] Once again, another researcher may prefer a different aggregation mode, such as a minimum. We found that an average worked the best for our context.

[14] Preregistration available at https://osf.io/xyab3/overview



Figure 3: Example lottery choice task from our experiment.

4.2 Theory Separation and Empirical Results

A design criterion is useful insofar as better designs according to the criterion produce stronger statistical results. We assess the extent to which this is true on our designs for a range of different target theories. We focus on two statistics in particular: the log likelihood ratio (LLR) the Vuong z-statistic for model selection, which we referred to as $\hat{D}$ and $\hat{z}$ respectively in Section 2.3. To compute $\hat{D}$ and $\hat{z}$ on a given design for model classes M_1, M_2, we use maximum likelihood to estimate $\hat{\theta}_1$ and $\hat{\theta}_2$ on the data from that design and then follow Equation 8. For both statistics, a positive value indicates that Theory 1 fits the data better; a negative value indicates that Theory 2 fits the data better; and a larger absolute value indicates that the winning theory has a larger advantage over the losing theory.[15]

Guided by Corollary 1, we look at the relationship between $\bar{\rho}$ and $|\hat{z}|$ across all 60 experiments for each of our eight theory pairs. Specifically, for each pair, we calculate ex ante $\bar{\rho}$ on all 60 designs according to our priors. For nested theories, we take $\bar{\rho}$ to be the one-sided criterion corresponding to the larger theory. For non-nested theory pairs, we weight both theories equally, i.e.

$$\bar{\rho} = 0.5\bar{\rho}_1 + 0.5\bar{\rho}_2.$$

Then, we compute the standard-normal statistic $\hat{z}$ for the same pair using the observed data. $|\hat{z}|$ is our measure of the *strength* of the statistical result: the larger this value, the

[15] In Appendix F.4, we reproduce the $\hat{z}$ regressions using a non-degenerate alternative to the Vuong statistic proposed by Shi (2015). Relationships are virtually unchanged.

more decisively the winning model beats the losing model.

We regress $|\hat{z}|$ on $\sqrt{NT}\ \bar{\rho}$ and a constant for each theory pair. We multiply $\bar{\rho}$ by $\sqrt{NT}$ so as to put the ex-ante design criterion on the same scale as the ex-post statistic; but the value of NT is approximately the same across designs, so virtually all variation comes from $\bar{\rho}$ itself. Recall from Corollary 1 that if data were actually drawn according to the hierarchical prior, we would expect the coefficient to be 1 for each theory pair. We of course do not believe that the prior literally describes the data generating process; but a *positive* coefficient indicates that designs with higher $\bar{\rho}$ are predictably producing stronger results as desired.

In Table 3, Column 1, report the coefficients from these regressions. We see that the coefficients are indeed positive for every single theory pair. The strength of the relationship varies across different theory pairs, with 4 of 8 being significant at $p = 0.05$. To give a sense of magnitudes, the coefficient on $\bar{\rho}_{CPT}$ in row 1 implies that moving from the median design ($\bar{\rho} = 0.04$) to the best design ($\bar{\rho} = 0.18$) increases the value of the model selection z-statistic from 1.39 to 2.85 at our sample size ($T = 8, N \approx 30$). In Column 2, we run analogous regressions using $\bar{D}$, the prior-averaged separation, and it's empirical analog $\hat{D}$, the average log likelihood ratio defined in Equation 8. The results are similar.

In addition to these comprehensive exercises, which allow us to separately assess the usefulness of the criterion for each theory-pair on all available data, we ran a simpler, pre-registered aggregated analysis comparing the absolute log likelihood ratio on optimal designs versus alternative designs. Results are qualitatively similar and significant at $p = 0.01$. A full description of this analysis can be found in Appendix F.3.

4.3 Separation in the Literature

We see in our experimental data that designs with higher separation tend to produce stronger statistical results. Now, we want to examine how our algorithmically-generated designs compare to what researchers are already doing. We focus on choice under risk precisely because this is a highly-studied experimental paradigm. To the extent that there are blind spots even in a literature like this, we are optimistic that our algorithm will be broadly useful for generating novel designs. We assemble a first-of-its-kind dataset of 146 experimental designs in choice under risk, drawn from 95 unique papers and 38 years of research.

4.3.1 Dataset Assembly

There are many different experimental paradigms used to study risky choice behavior, but we narrow our focus to binary choice designs where subjects make at least two separate

Theory pair	$\lvert\hat{z}\rvert \sim \sqrt{NT}\,\bar{\rho}$		$\lvert\hat{D}\rvert \sim \bar{D}$	
CPT vs. EU	0.69***	(0.23)	1.08***	(0.23)
RCPT vs. REU	0.48***	(0.09)	0.40***	(0.07)
CC vs. CPT	0.49***	(0.11)	0.74***	(0.10)
CC vs. RCPT	0.27	(0.17)	0.34**	(0.14)
CC vs. SAL	0.68***	(0.14)	0.80***	(0.12)
CPT vs. RCPT	0.08	(0.19)	0.33*	(0.17)
CPT vs. SAL	0.44*	(0.26)	0.07	(0.16)
RCPT vs. SAL	0.08	(0.25)	0.37*	(0.22)

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

Table 3: Columns (1) and (2) report regression coefficients (s.e.) from regressing the empirical absolute Vuong statistic and absolute log likelihood ratio on normalized separation $\bar{\rho}$ and separation $\bar{D}$ respectively. All regressions include a constant term. Note that N is held roughly fixed across designs, so virtually all variation in the column (1) come from $\bar{\rho}$.

choices between a pair of lotteries. This means we exclude valuation and choice list designs, even if technically a choice list is implemented through a table of binary choices.[16] To a first approximation, binary choice designs share a domain with our own experiments, which allows us make apples-to-apples design power comparisons. We exclude designs where probabilities or payoffs are not specified (as in Ellsberg experiments), and we also exclude designs where payoffs cannot be clearly ordered (e.g., some lotteries pay out apples and others pay out oranges).

We begin by compiling a list of candidate articles using the EconLit database. We pulled all journal articles satisfying one of the two following conditions:

1. JEL code d81 and one of codes {d91, c90, c91}. Code d81 is for "Criteria for Decision-Making under Risk and Uncertainty," d91 is for "Micro-Based Behavioral Economics," and c91 and c90 are both related to experiments.

2. Abstract contains a term related to risky choice and a term related to experiments. Risk terms: {gamble, risk preference, risk preferences, choice under risk, expected utility, lottery}. Experiment terms: {experiment, experiments, experimental, experimentally}.[17]

[16]We consider choice lists to be a separate form of behavioral elicitation because binary choices in sequential price list rows are not plausibly independent, as is assumed in our theory.

[17]The reason we did not use "risk" on its own as a risk term is that it increased search results by an order of magnitude, and by inspection the additional articles seemed largely irrelevant to our interests.

We limit to articles published in one of the Top 100 Economics journals, according to the IDEAS/RePEc aggregate rankings in December of 2025. This search and filter procedure gives 1,217 candidate articles. We download PDFs for all of these articles and extract the first 8 pages of text. We use a large language model to scan this text and flag articles which discuss an experiment with binary lottery choices. 233 articles are flagged, and we go through each of these by hand to extract the experiment. In total, 95 articles include an extractable experimental design that meets our criteria.[18]

For each design, we record the probabilities and payoffs of both lotteries for each binary choice task. If the lotteries are communicated to participants in a common state space and the correlation structure is described in the text, this is recorded. Any compound structure is also recorded, though none of the theories we study here make explicit use of compoundness. If a paper contains distinct experiments which are analyzed separately, these are recorded as separate designs. If an experiment includes some tasks that meet our criteria and others that do not (e.g., some binary menus and some trinary menus), we include only the tasks that meet the criteria and record it as partial design. We do not record the payoff units, only numerical quantities as they are presented to participants.

4.3.2 Analyzing Separation

We can compute $\bar{\rho}$ on all of these designs for any theories we like. However, we focus on separation of EU and CPT because these are the two most widely discussed theories in this literature. If our algorithmic designs meaningfully improve upon the literature in terms of $\bar{\rho}_{CPT}$, which aligns with the explicit empirical goals in many of these papers, we take this as strong support for an algorithmic approach to design more generally.

In order to calculate and summarize $\bar{\rho}$, we must make a consequential choice on how to handle the fact that the scale of payoffs varies widely across experimental designs – a classic Allais paradox problem has a potential \$5,000,000 payout, while other experiments may have a max payout of \$0.50. Taken literally, any stochastic theory with Fechnarian noise (as in our formulation of CPT and EU) would predict very different choice probabilities on these two designs. Any noise level which generates meaningful variation in choice probabilities on the \$0.50 design will imply virtually deterministic choice on the \$5,000,000 design. However, no reasonable analyst would actually expect a parameterized model to extrapolate

[18]The most common reasons for excluding papers were ambiguous gambles as in Ellsberg experiments, re-analyses rather than new experiments, dynamically defined choices as in staircase elicitations, and theory-only papers. There were also several experiments which met our criteria, but in which the set of lottery tasks was not explicitly given in the paper or appendix.

in this way. The noise is not the substantive part of these models; the preference parameters are. Since none of our theories are substantively about how choice changes with scale, we opt to standardize the payoff-level of all designs before calculating power. Specifically, for each design, we first divide all payoffs by $M/50$, where M is the maximum absolute payoff in the whole design. This puts each literature design on exactly the same payoff scale as our own experiments.

Another source of variation across these designs is the number of tasks. Intuitively, more tasks in a design should always increase expected model separation, since it generates more data.[19] However, recall that our unit of analysis $\bar{\rho}$ refers to the *average* power contributed by each task in the design. Practically, $\bar{\rho}$ gives us an even playing field to compare designs of different sizes; but there are also two reasons why we might actually care more about this object than total power. The first reason is efficiency – if a design with many tasks has small $\bar{\rho}$, then even if there are sufficiently many tasks to separate theories, a researcher could save resources by running a smaller design with larger $\bar{\rho}$. The second reason has to do with *why* we run experimental tests of theory at all. A good experiment should illustrate clearly to an analyst how a theory fails and why. In our view, it is easier to draw insights from an experiment where a theory is far from the truth on average, as opposed to a large collection of small failures.

We begin by summarizing design power to separate CPT from Expected Utility. In Figure 4, we plot the literature's distribution of power for separating CPT from EU, splitting designs into subgroups based on whether they discuss testing EU and CPT.[20] This figure illustrates several facts. First, designs explicitly focused on EU and CPT indeed have higher power to separate CPT from EU (mean $\bar{\rho}$ of 0.059 vs. 0.040, $p = 0.01$). This indicates to us that researchers in this literature are generally aligned with our proposed objective, and have a real sense of which design features make experiments more powerful. Second, we see that our algorithmic designs ($\bar{\rho} = 0.18$) far outpace even the best literature designs in terms of power ($\bar{\rho} = 0.12$). In our experiment, a change in $\bar{\rho}$ of this size for CPT-EU was associated with an increase of 0.65 in the z-statistic for selecting CPT over EU. This suggests that the current approach to design is leaving power on the table, which our algorithmic procedure could recoup.

It is impossible to say definitively that existing designs are suboptimal given the researcher's goals, which we cannot observe. Many papers are not intended to separate EU

[19] Strictly speaking, more tasks need not increase the experiment's power to correctly select the true model, if new tasks introduce sufficiently high variance.

[20] We record by hand whether an experiment is doing some form of theory comparison or testing; and we also record which theories are discussed.

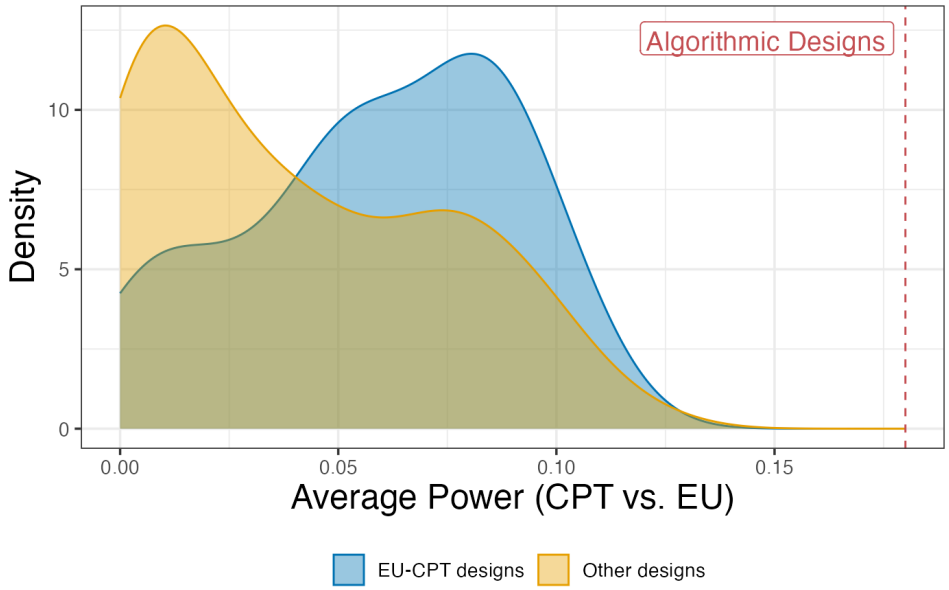
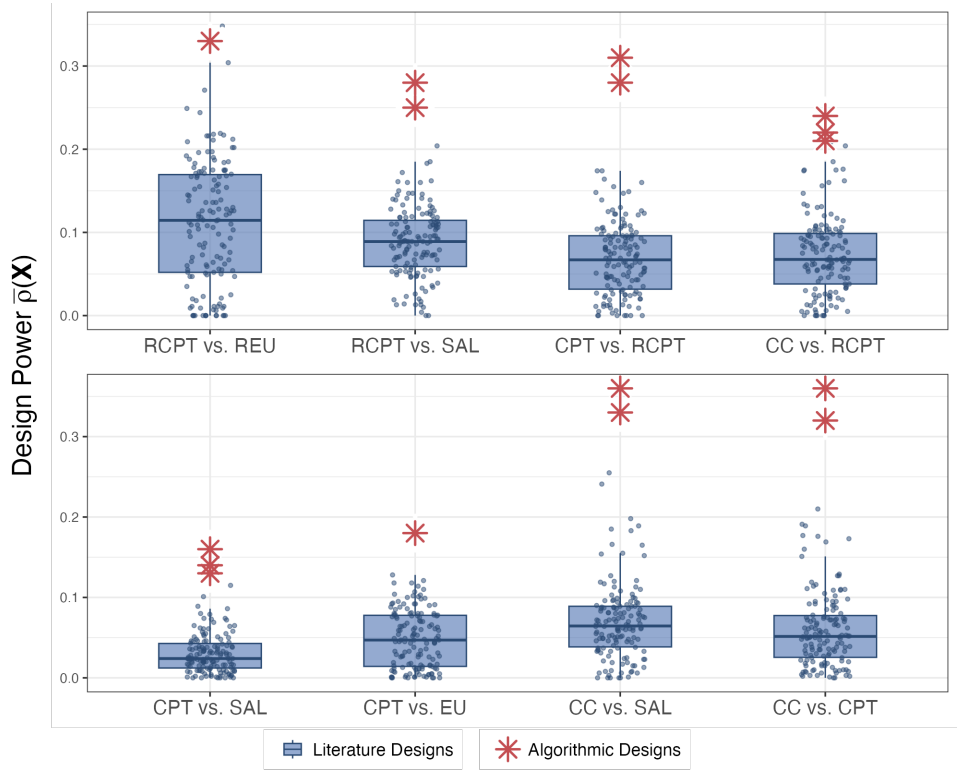


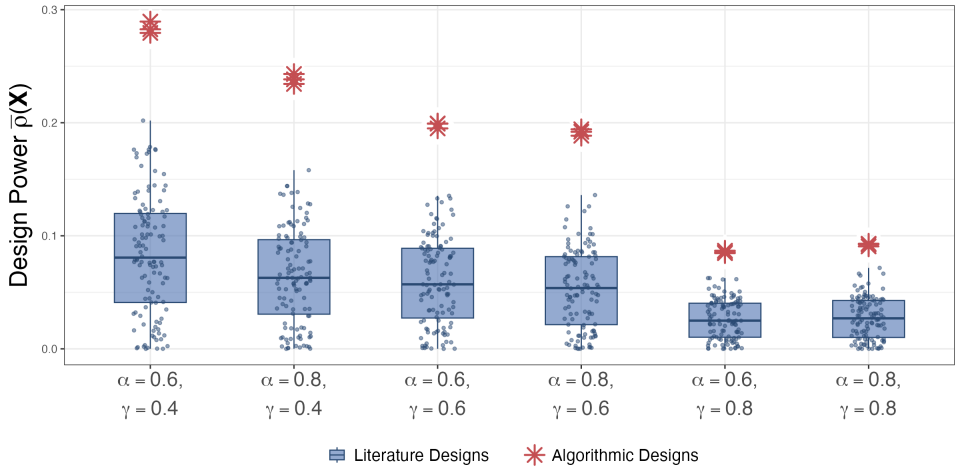
Figure 4: Average Power to Separate CPT from EU. The EU-CPT design classification includes designs which are both (1) doing some form of theory testing and (2) explicitly discuss both EU and Rank-Dependent Utility (a class of models which includes CPT).

from CPT at all, and even those that are may have different beliefs about levels of risk aversion or probability distortion. We conduct two exercises which cast doubt on the possibility that these facts drive apparent sub-optimality. First, we inspect the design power distribution for the 7 alternative theory pairs from our experiment, comparing each to the design power of algorithmic designs optimized for that pair. For every single pair except for RCPT-REU, we observe a substantial gap between the best-powered literature design and the best-powered algorithmic design (see Figure 5a).

Second, we inspect the design power distribution compared to algorithmic design power under many possible parameter values for CPT. Recall that when we calculate $\bar{\rho}$, we are taking an average over a set of hypothesized parameter values determined by us. In principle, it could be the case that our algorithmic design does well on average; but worse for some particular parameter values. To assess this, we can separately calculate $\bar{\rho}$ for different hypothesized values of α and γ, both in the literature and on our algorithmic design. We report results in Table 5b. We see variation in the overall level of $\bar{\rho}$ depending on which parameter values we use – in particular, $\bar{\rho}$ is lower when γ is closer to 1, which reflects that less probability distortion make CPT more like EU. Despite these level differences, design power of the algorithmic designs is consistently higher than even the best existing design. We also find this reassuring because it indicates that even if researchers don't select quite



(a) Average normalized separation for alternative theory pairs.



(b) CPT-EUT normalized separation for different CPT parameters.

Figure 5: Separation of literature designs and best algorithmic designs for different theory pairs (Panel a) and priors (Panel b)

the right parameter values in the power objective, there are still benefits to optimizing.

5 Separation and Experimental Results

Thus far, we have focused on how the normalized separation criterion can be deployed to help researchers design more powerful experiments. In this section, we explore how the same machinery can illuminate ex-post how an experiment's results may be driven by design choices. We also revisit a classical experimental design and discuss how the separation criterion can be used to suggest new interpretations of existing results.

5.1 Predicting Results on Asymmetric Designs

When testing between two non-nested theories, we typically want a design on which Theory 2 would fit poorly if Theory 1 were true (high $\bar{\rho}_1$); and Theory 1 would fit the data poorly if Theory 2 were true (high $\bar{\rho}_2$). We've thus suggested a natural criterion is

$$\bar{\rho} \equiv 0.5\bar{\rho}_1 + 0.5\bar{\rho}_2.$$

However, when we do not approach design systematically, there is no guarantee that $\bar{\rho}_1$ and $\bar{\rho}_2$ will move together. Instead, we may inadvertently produce *asymmetric* designs: designs on which $\bar{\rho}_1$ is high while $\bar{\rho}_2$ is low, or vice versa.

Asymmetry has important consequences for experimental results. Consider a design with high $\bar{\rho}_1$, low $\bar{\rho}_2$. This means that if Theory 1 generates the data, Theory 2 will not be able to explain the results; but if Theory 2 generates the data, Theory 1 will still be able to explain the results well. Regardless of which theory generates the data, we will never see Theory 1 fitting poorly – at worst, it should have a similar fit to Theory 2. Put differently, it's as though on this design, Theory 1 approximately nests Theory 2, even though the two theories are, strictly speaking, not nested. We say that a design like this *asymmetrically favors* Theory 1.

To see why asymmetry can complicate how we interpret an experiment's results, it's useful to consider two hypothetical designs: Design 1, which asymmetrically favors Theory 1; and Design 2, which asymmetrically favors Theory 2. If we believe that one of these theories is correct, say Theory 1, then asymmetry is simply a power problem. Design 1 is well-powered to show us that Theory 1 beats Theory 2; and Design 2 is not well-powered. If instead we believe that both theories may be misspecified, then we have a different problem.

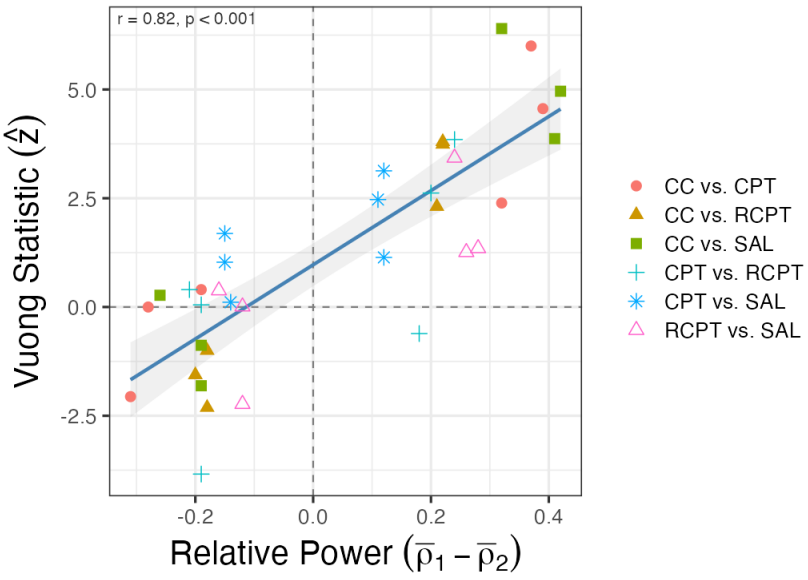


Figure 6: Experimental Results on Asymmetric Designs. Each point is an individual design. The x-axis is the difference $\bar{\rho}_1 - \bar{\rho}_2$, which is an ex-ante measure of asymmetry: large positive values indicate a design asymmetrically favors Theory 1, and large negative values indicate a design asymmetrically favors Theory 2. The y-axis is a standard normally distributed test statistic constructed from observed data. A positive value indicates that Theory 1 fits the data better, a negative value indicates that Theory 2 fits the data better, and the threshold for statistical significance at $p = 0.05$ is $|z| > 1.96$.

When we run these two experiments, we may well find that Theory 1 statistically significantly fits the data better on Design 1; and Theory 2 statistically significantly fits the data better on Design 2. Inconsistency in and of itself is not necessarily problematic – if it's the case that one theory is more appropriate in some parts of the domain than others, this may be interesting for researchers to learn. In this particular case, though, the inconsistency is a predictable consequence of the choice of design. If we can predict before collecting any data that Theory 1 will fit better on Design 1, and Theory 2 will fit better on Design 2, simply because of the designs' construction; then we should be cautious in interpreting what observing data in these experiments teach us about the theories.

We assess the scope for such issues using a subset of our lottery choice designs, which were specifically constructed to produce asymmetry (see Section 4.1.1). For each of 6 non-nested theory pairs in Table 2, we generated 3 designs which asymmetrically favor Theory 1, and 3 designs which asymmetrically favor Theory 2. We ran each of these 36 designs and computed the Vuong z-statistic for the corresponding pair of theories. Results are presented visually in Figure 6.

On the left side of the figure, we have the designs which asymmetrically favor Theory 2; and on the right side, we have the designs which favor Theory 1. The higher the value of

$\hat{z}$ on the y-axis, the more the experimental results favor Theory 1. The first thing we notice is that, for all 6 theory pairs, we get higher values of $\hat{z}$ on the designs which favor Theory 1 vs. the designs which favor Theory 2. But even more striking, this isn't a pure power effect, where the design choice simply moderates the margin by which the winning theory wins – we often observe *reversals* in which theory fits the data better as we move from one set of designs to the other. For some theory pairs, we are even getting reversals of statistically significant outcomes (i.e., $\hat{z} < -1.96$ for designs favoring Theory 2, and $\hat{z} > 1.96$ for designs favoring Theory 1). In Appendix F.3, we do the same exercise using the raw log likelihood ratio as the outcome and results are virtually unchanged.[21]

In addition to these targeted comparisons, we can also look at how results vary as a function of asymmetry across all 60 designs for each theory pair. Specifically, we can run regressions analogous to those in Table 3, except now we regress $\hat{z}$ on $\sqrt{NT}\bar{\rho}_{\Delta}$, where $\bar{\rho}_{\Delta} \equiv \bar{\rho}_1 - \bar{\rho}_2$. This exercise is also theoretically motivated by Corollary 1. We do this and report the results in Table 4, Column 1. Much like our reduced form exercise in Figure 6, positive coefficients in these regressions indicate that as ex-ante power tips in favor of Theory 1 across all 60 designs, ex-post statistical evidence tips in favor of Theory 1 as well. We see that the coefficients are indeed positive for all theory pairs, and significant at $p = 0.05$ for 5 of 6.[22]

Theory pair	$\hat{z} \sim \sqrt{NT}\,\bar{\rho}_{\Delta}$		$\hat{D} \sim \bar{D}_{\Delta}$	
CC vs. CPT	1.25***	(0.20)	1.44***	(0.16)
CC vs. RCPT	0.86***	(0.24)	0.55***	(0.18)
CC vs. SAL	1.36***	(0.16)	1.21***	(0.14)
CPT vs. RCPT	1.01**	(0.43)	1.24***	(0.38)
CPT vs. SAL	0.56	(0.36)	0.06	(0.21)
RCPT vs. SAL	1.19**	(0.46)	0.96***	(0.35)

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

Table 4: Columns (1) and (2) report regression coefficients (s.e.) from regressing the empirical Vuong statistic and the log likelihood ratio on the theory-pair difference normalized separation ($\bar{\rho}_{\Delta} = \bar{\rho}_1 - \bar{\rho}_2$) and separation $\bar{D}$ ($\bar{D}_{\Delta} = \bar{D}_1 - \bar{D}_2$) respectively. All regressions include a constant term. Normalized separation is scaled by the sample size NT, but N is held roughly fixed across designs, so virtually all variation in the column (1) come from $\bar{\rho}_{\Delta}$.

[21]We pre-registered the log likelihood ratio version of this exercise, but we prefer the version in Figure 6 because the units of the outcome variable are more interpretable.

[22]Note that we do not include the nested pairs for this exercise, since there would be no way for the smaller theory to be asymmetrically powered against the larger theory under our definitions.

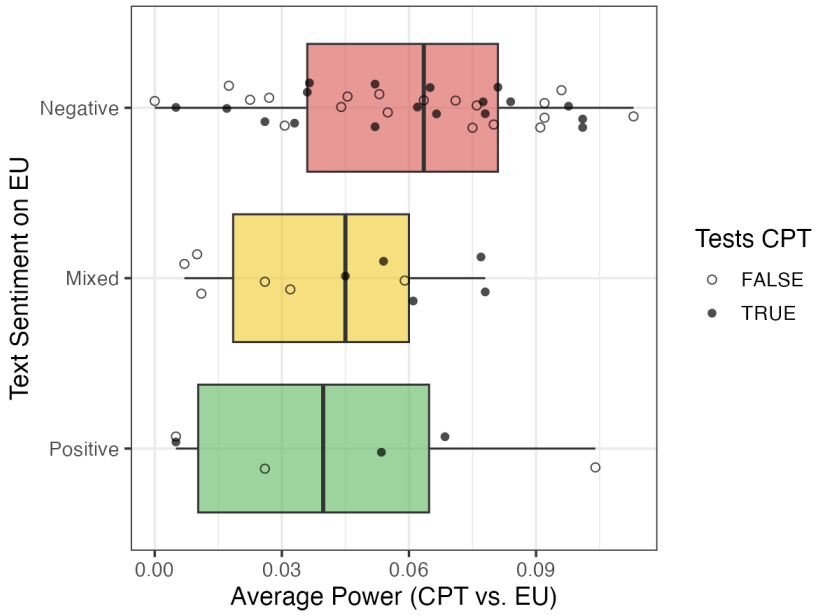


Figure 7: Design power for separating CPT from EUT, grouped by paper's EUT sentiment.

5.2 Literature Results and Ex-Ante Power

Given our empirical exercise in the previous section, a natural question is whether design power can help us understand any variation in results across different experimental work. Conducting the same exercise here is challenging because of the wide variation in model specifications and reported statistics across all of these papers. Some papers report log likelihoods, some report mean squared errors, and still others report rates of axiom violation.

Given these issues, we attempt a less formal exercise based on the text of the papers. First, we limit to the 60 papers which we classified as testing expected utility theory or an axiom of expected utility theory. Then, we prompt Claude Opus 5.5 to read the text of these papers and characterize whether the sentiment towards Expected Utility Theory is Positive, Negative, or Mixed. We exclude six papers with no clear sentiment. In Figure 7, we compare design power for separating CPT from EUT in each of the three groups.

If CPT were generally the reason EUT fails to explain data, we would expect to see that as design power increases, we tend to get results that are *more negative* towards EUT. The relationship appears to go in this direction, but the correlation is weak.[23] Still, considering these designs through the lens of power can help us interpret experimental results.

We make two observations. First, we notice that the papers which came to a positive conclusion about EUT tend to be low-powered (most are around 1/3 the value of our algo-

[23]A difference in means test of the "Negative" vs. "Mixed/Positive" papers has $p = 0.07$.

Choice 1

Option A: $5 million with probability 10%, $1 million with probability 89%, $0 otherwise
Option B: $1 million with probability 100%

Choice 2

Option C: $5 million with probability 10%, $0 otherwise
Option D: $1 million with probability 11%, $0 otherwise

Figure 8: Allais's common consequence paradigm. The Allais paradox refers to the modal choice pattern of selecting Option B in Choice 1 and Option C in Choice 2, which violates expected utility.

rithmic design). This should make us take the success of EUT on these designs with a big grain of salt – even if CPT were right, we would expect EUT to fit the data well. Second, we notice that many *very* low-powered designs deliver negative evidence for EUT. If our CPT-prior were the right alternative description of behavior, this wouldn't happen – so what does this tell us? It tells us exactly that the violations of EUT observed on these designs are not what we would predict under CPT. Often this is apparent reading the papers: many don't claim to test or study CPT at all, but instead focus on other non-EUT alternatives, or simply document a failure. Another way of putting this is that on many of these designs, the behavioral patterns that violate EUT would also violate CPT.

We can also learn something from the papers which explicitly study CPT, yet reject EUT with a low-powered design. To illustrate, we focus on one particular highly-studied lottery choice design: Allais's common consequence problem (Allais, 1953). The design is composed of the two binary choices in Figure 8.

The design is famous because it canonically illustrates a failure of EUT; under EUT preferences, a rational subject should either select A in Choice 1, C in Choice 2; or B in Choice 1, D in Choice 2. Instead, subjects often select B in Choice 1 and D in Choice 2. This paradox has been generative for theorists – many theories of non-EU risk preferences explicitly use it as a motivating example, including the seminal Tversky and Kahneman (1992) (KT) which first introduced Cumulative Prospect Theory.

If we return to Figure 4, we notice something puzzling: the classic Allais paradox, directly cited in Tversky and Kahneman (1992), is a very low-powered design for separating CPT from EU. How can this be true, when CPT was constructed precisely to explain a known breakage of EU on this pair of problems? It turns out that at all of the CPT parameter values we considered, guided by the later literature, CPT does *not* predict a breakage of EU. You

might worry that this is hindsight bias – but we can also use the CPT parameters estimated by KT in this very paper. At these estimates ($\alpha = 0.88$, $\gamma = 0.61$), KT's own model predicts a subject would prefer the risk-averse option in both choices.[24]

Concretely, this means that CPT–as we know it–actually *does not* explain EU violations on the Allais Paradox. We are not the first to make this observation – the particular point about Allais was also made by Neilson and Stowe (2002), and McGranaghan et al. (2026) show that CPT fails to explain behavior on a broader class of common consequence problems. The disconnect comes from separately considering what the theory *can* accommodate, and what parameter values actually explain the data. Conceptually, CPT explains the failure of the independence axiom by positing that subjects overweight the 1% state in Choice 1. In order to actually get a preference flip, though, the degree of overweighting needs to be very large – much larger than the level we observe when we look at the collection of choices in KT's experiment, or elsewhere in the risky choice literature. The incongruence here is obvious once we calculate power using the estimated parameters, but goes unmentioned in the paper and likely unnoticed by the authors.

Zooming back out, these examples illustrate two broad types of inference errors that our framework for design power can help us avoid. First, we should be careful about interpreting "positive" evidence for a theory on a design which is low-powered to separate it from alternatives. And second, if we find negative evidence for a theory on a design we expected to be low-power relative to some alternative; this tells us we're likely thinking of the wrong alternative.

6 Conclusion

In this paper, we have developed a framework for systematically approaching experimental design. We proposed a quantitative criterion for design based on statistical power for model selection and laid out an algorithmic procedure for generating highly-powered designs. The criterion and design procedure are general enough to be suitable for a wide range of researcher goals. In the paper we have focused on lottery experiments and theories of choice under risk; but the framework is applicable for design in any setting where the theories' predictions depend on design parameters. The experimental game theory literature is one natural application for our procedure, as are product choice experiments with quantitative models of context dependence.

[24]KT use power utility and a probability-weighting function of the form $w(p) = \frac{p^\gamma}{(p^\gamma+(1-p)^\gamma)^{1/\gamma}}$.

After developing this machinery for designing experiments, we demonstrated its usefulness in lottery choice experiments. We ran a set of 60 new experimental designs and showed that more powerful designs according to the criterion tended to produce stronger statistical results. Furthermore, we conducted a meta-analysis of the existing choice-under-risk literature which indicated that standard practices typically do leave statistical power on the table.

Finally, we explored using our own experiments and a further analysis of the literature how a formal design criterion not only allows us to design better experiments; but gives us a new way of understanding how features of a design drive experimental results. We showed experimentally that asymmetries in design power can translate into predictable variation in statistical results, shifting them in favor of the asymmetrically-powered theory. In the existing literature, we saw suggestive evidence that some null or weak results likely reflect designs which were simply underpowered to show anything else. And finally, we discussed that when an apparently "low-powered" actually does come to a strong statistical conclusion, this should either make us revisit our priors on theory parameters, or question whether results reflect another explanation than our considered models.

References

Allais, M. (1953). Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l'Ecole Americaine. *Econometrica*, 21(4):503–546.

Atkinson, A. C. and Fedorov, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70.

Balietti, S., Klein, B., and Riedl, C. (2021). Optimal design of experiments to identify latent behavioral types. *Experimental Economics*, 24(3):772–799.

Barseghyan, L., Coughlin, M., Molinari, F., and Teitelbaum, J. C. (2021). Heterogeneous Choice Sets and Preferences. *Econometrica*, 89(5):2015–2048. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17448.

Bland, J. R. (2025). Optimizing Experiment Design for Estimating Parametric Models in Economic Experiments.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience Theory of Choice Under Risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Bronars, S. G. (1987). The Power of Nonparametric Tests of Preference Maximization. *Econometrica*, 55(3):693–698.

Camerer, C. F. and Ho, T.-H. (1994). Violations of the betweenness axiom and nonlinearity in probability. *Journal of Risk and Uncertainty*, 8(2):167–196.

Castillo, M., Jordan, J. L., and Petrie, R. (2018). Children's rationality, risk attitudes and field behavior. *European Economic Review*, 102:62–81.

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., and Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, 47(3):255–289.

Chapman, J., Snowberg, E., Wang, S., and Camerer, C. (2024). Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters.

Choi, S., Fisman, R., Gale, D., and Kariv, S. (2007). Consistency and Heterogeneity of Individual Behavior under Uncertainty. *American Economic Review*, 97(5):1921–1938.

Dette, H., Melas, V. B., and Guchenko, R. (2015). Bayesian T-optimal discriminating designs. *The Annals of Statistics*, 43(5).

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). Scalable Global Optimization via Local Bayesian Optimization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fudenberg, D., Gao, W., and Liang, A. (2023). How Flexible Is That Functional Form? Quantifying the Restrictiveness of Theories. *The Review of Economics and Statistics*, pages 1–50.

Fudenberg, D. and Liang, A. (2019). Predicting and Understanding Initial Play. *American Economic Review*, 109(12):4112–4141.

Healy, P. J. and Leo, G. (2024). Minimal Experiments.

Kahneman, D. and Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica (pre-1986)*, 47(2):263–291. Num Pages: 29.

López-Fidalgo, J., Tommasi, C., and Trandafir, P. C. (2007). An Optimal Experimental Design Criterion for Discriminating Between Non-Normal Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):231–242.

McGranaghan, C., Nielsen, K., O'Donoghue, T., Somerville, J., and Sprenger, C. D. (2026). Connecting Common Ratio and Common Consequence Preferences. *Journal of Political Economy*.

Mullainathan, S. and Rambachan, A. (2023). From Predictive Algorithms to Automatic Generation of Anomalies.

Myung, J. I. and Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review*, 116(3):499–518. Place: US.

Neilson, W. and Stowe, J. (2002). A Further Examination of Cumulative Prospect Theory Parameterizations. *Journal of Risk and Uncertainty*, 24(1):31–46.

Niederle, M. (2025). Experiments: Why, How, and A Users Guide for Producers as well as Consumers.

Ponce De Leon, A. C. and Atkinson, A. C. (1991). Optimum experimental design for discriminating between two rival models in the presence of prior information. *Biometrika*, 78(3):601–608.

Prelec, D. (1998). The Probability Weighting Function. *Econometrica*, 66(3):497–527.

Schennach, S. M. and Wilhelm, D. (2017). A Simple Parametric Model Selection Test. *Journal of the American Statistical Association*, 112(520):1663–1674. _eprint: https://doi.org/10.1080/01621459.2016.1224716.

Shi, X. (2015). A nondegenerate Vuong test. *Quantitative Economics*, 6(1):85–121.

Shubatt, C. and Yang, J. (2024). Tradeoffs and Comparison Complexity. arXiv:2401.17578 [econ.GN].

Snowberg, E. and Yariv, L. (2025). Evaluating experimental designs. In Snowberg, E. and Yariv, L., editors, *Handbook of Experimental Methodology*, volume 1 of *Handbook of Experimental Methodology*, pages 1–24. North-Holland.

Starmer, C. and Sugden, R. (1991). Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation. *The American Economic Review*, 81(4):971–978.

Tommasi, C. and López-Fidalgo, J. (2010). Bayesian optimum designs for discriminating between models with any distribution. *Computational Statistics & Data Analysis*, 54(1):143–150.

Tversky, A. and Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk & Uncertainty*, 5(4):297–323.

van der Vaart, A. (1998). Asymptotic Statistics (1998). *eTextbooks for Students*.

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2):307–333.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25.

A General Decision Theory

A.1 The General Decision Problem

The model-selection problem in the main text is one instance of a more general experimental design problem. Let $X \in \mathcal{X}_\ell^T$ be a feasible design and let $\mathbf{Y}$ denote the data it generates. The data-generating process is an unknown element p_0 of a set $\mathcal{P}$; a researcher may summarize prior knowledge about p_0 with a probability measure Π on $\mathcal{P}$. After observing the data, the researcher takes an action a from an action space $\mathcal{A}$ using a decision rule

$$\delta : \mathcal{Y}^{NT} \longrightarrow \mathcal{A}.$$

The loss from taking action a when the truth is p_0 is $\ell(a, p_0)$. Holding the decision rule fixed, the ex ante loss of design X is therefore

$$\mathcal{L}(X;\delta) = \int_{\mathcal{P}} \int \ell(\delta(\mathbf{Y}), p_0)\; p_0(d\mathbf{Y} \mid X)\, \Pi(dp_0). \tag{10}$$

The general design problem is to choose

$$X^* \in \arg\min_{X \in \mathcal{X}_\ell^T} \mathcal{L}(X;\delta).$$

The researcher must choose X before observing $\mathbf{Y}$, so this choice necessarily uses prior knowledge about which data-generating processes are reasonable candidates for p_0. We let $\mathcal{P}$ denote the candidate models, and Π the weights the researcher assigns to each model in $\mathcal{P}$. We do not need to interpret Π as a Bayesian prior, or assume that the true model is literally drawn from $\mathcal{P}$ according to Π. Instead, we think of $\mathcal{P}$ and Π as a convenient way to translate approximate knowledge about the true data generating process into a concrete design criterion. Note that these objects are used to formulate our scientific goal (the loss function); but play no role in the statistical procedure used after data collection (the decision rule).

A.2 Model Selection

Suppose the researcher compares K model classes $\mathcal{M}_k = \{g_k(\cdot \mid \cdot, \theta_k) : \theta_k \in \Theta_k\}$. In this case, a natural candidate set is $\mathcal{P} = \bigcup_{k=1}^K \mathcal{M}_k$. The prior knowledge represented by Π can be read as a hierarchical Bayesian prior. First draw a model index k with probability P_k,

then draw θ_k from a distribution π_k over Θ_k, and set $p_0(\cdot \mid X_t) = g_k(\cdot \mid X_t, \theta_k)$. The action is the selected model $\hat{k}$, and the loss $\ell(a, p_0) = -\mathbf{1}\{a = k\}$ rewards correct selection. In this special case, minimizing (10) is equivalent to maximizing the weighted-average probability of correct selection.

$$X^* \in \arg\max_{X \in \mathcal{X}_\ell^T} \sum_{k=1}^{K} P_k \int_{\Theta_k} \Pr\big(\hat{k} = k \mid X, (k, \theta_k)\big)\, \pi_k(d\theta_k).$$

The main text uses the same logic with a finite grid of plausible parameter values and researcher-specified weights.

A.3 GARP and Bronars' Power

Consider an experiment on revealed preference. A design $X = (X_1, \ldots, X_T)$ specifies the sequence of budget lines. Subject i chooses a bundle on each line, producing the record $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})$. Let $\mathcal{G}(X)$ be the set of records that satisfy the Generalized Axiom of Revealed Preference (GARP), so no revealed-preference cycle may contain a strict link. A test of GARP rejects when at least one observed record lies outside this set,

$$\delta(\mathbf{Y}_1, \ldots, \mathbf{Y}_N) = \mathbf{1}\{\exists i \le N : \mathbf{Y}_i \notin \mathcal{G}(X)\}.$$

To evaluate a design before collecting data, Bronars considers a simple benchmark data-generating process. For each candidate design X, let P_X^{U} denote the process under which every subject and every task draws a bundle uniformly along the corresponding budget line, independently across subjects and tasks. We can represent this benchmark in the general framework by taking $\mathcal{P}$ to contain these random-choice processes and by putting all weight on the uniform random-choice rule. This rule induces P_X^{U} for each candidate design X. Unlike the model-selection example, this is best understood as a convenient benchmark rather than a systematic Bayesian prior. Uniform random choice is a reference distribution used to compare designs, not a substantive claim about how preferences are generated. Bronars' power is the probability that the test rejects under this benchmark.

$$X^* \in \arg\max_{X \in \mathcal{X}_\ell^T} \Pr_{P_X^{\mathrm{U}}}(\exists i \le N : \mathbf{Y}_i \notin \mathcal{G}(X)). \tag{11}$$

Thus this example has the same structure as (10). The researcher specifies the action, the loss implicit in the testing goal, and a data-generating distribution that serves as the ex ante reference, then chooses the design with the best expected performance.

B Heterogeneity

Section 2 discusses the role of design power for model selection under the assumption that subjects are homogeneous. More broadly, researchers are often interested in heterogeneity across subjects and use experiments not only to select one model, but also to estimate the population share of people whose behavior is consistent with each model. In this section, we analyze common procedures used to estimate population shares, and show that the effect size $\bar{\rho}$ and the resulting power $\Phi(\sqrt{T}\,\bar{\rho})$ also govern the statistical properties of the estimators.

We first describe the data-generating process implicitly assumed in many papers that analyze heterogeneity. We then study two procedures used in practice. Our goal here is not to match every empirical implementation exactly, but to capture the role of design power in a broad class of applications.

We assume the data generating process is as follows. There are N subjects and T tasks. Subject i has an unobserved type $Z_i \in \mathcal{K} = \{1, \ldots, K\}$, with population shares w^0, and a type-specific parameter $\theta_i \mid Z_i = k \sim F_k(\cdot \mid \eta_k^0)$. Conditional on $(Z_i = k, \theta_i)$, outcomes are independent across tasks with $Y_{it} \sim g_k(\cdot \mid X_t, \theta_i)$. The point-mass case $F_k = \delta_{\theta_k}$ is the finite mixture where subjects within a type share a common set of parameters, whereas a nondegenerate F_k allows parameters to vary within type.

The first procedure is *individual classification*. The researcher assigns each subject to a type, either by estimating the subject's best-fitting model or by applying a prespecified rule such as "the subject violates the model's prediction." She then counts the assignments to estimate type shares (Choi et al., 2007; Cavagnaro et al., 2013). The second procedure is *mixture estimation*, in which the researcher treats types as latent and estimates population shares jointly with the model parameters, typically by maximum likelihood (e.g., Barseghyan et al., 2021). For both procedures, the estimand is the population share w_k^0 of each type. We next characterize how design power affects these estimators.

B.1 Individual classification

To illustrate the first procedure formally, consider the case where the individual-level classification is based on likelihood comparisons.

$$\hat{\theta}_{k,i} \in \arg\max_{\theta \in \Theta_k} \sum_{t=1}^{T} \log g_k(Y_{it} \mid X_t, \theta),$$

$$\hat{Z}_i \in \arg\max_{k \in \mathcal{K}} \sum_{t=1}^{T} \log g_k(Y_{it} \mid X_t, \hat{\theta}_{k,i}), \qquad \hat{w}_k^{\text{cls}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{Z}_i = k\}.$$

For $K = 2$, define subject i's fitted average likelihood advantage of class k over the other class as

$$\hat{D}_{ik}^{\text{cls}} = \frac{1}{T} \sum_{t=1}^{T} \log \frac{g_k(Y_{it} \mid X_t, \hat{\theta}_{k,i})}{g_{3-k}(Y_{it} \mid X_t, \hat{\theta}_{3-k,i})}.$$

Define the confusion matrix by

$$Q_{kk'} = \Pr(\hat{Z}_i = k' \mid Z_i = k),$$

where the probability also averages over $\theta_i \sim F_k(\cdot \mid \eta_k^0)$. The law of large numbers gives the following basic result.

Proposition 2 (Individual classification)**.**
For any number of classes K and fixed T, as $N \to \infty$,

$$\hat{w}^{\text{cls}} \xrightarrow{p} Q^\top w^0.$$

For each class, writing $r = Q^\top w^0$, the share estimator has

$$\text{MSE}(\hat{w}_k^{\text{cls}}) = \frac{r_k(1-r_k)}{N} + (r_k - w_k^0)^2. \tag{12}$$

For $K = 2$, under analogous likelihood regularity and task-level Lindeberg conditions, as $T \to \infty$,

$$\frac{\sqrt{T}\left(\hat{D}_{ik}^{\text{cls}} - \bar{D}_{k,3-k}(\theta; X)\right)}{\sqrt{\bar{V}_{k,3-k}(\theta; X)}} \xrightarrow{d} \mathcal{N}(0,1)$$

conditional on $Z_i = k$ and $\theta_i = \theta$. Consequently, the Gaussian approximation to the diagonal

classification probabilities is

$$Q_{kk} \approx \mathbb{E}_{\theta\sim F_k}\left[\Phi\big(\sqrt{T}\,\bar{\rho}_{k,3-k}(\theta;X)\big)\right], \qquad Q_{k,3-k} = 1 - Q_{kk}. \tag{13}$$

Proposition 2 indicates that the population share estimator $\hat{w}^{\text{cls}}$ converges not to the true shares of types, but instead to the shares of assigned labels. If the individual-level classification is imperfect, then these two shares do not coincide. The confusion matrix characterizes the quality of individual-level classification, and its diagonal entries denote the probability of correct classification for each type. The two shares are close only when the classification matrix is close to the identity. In the two-class case, the Gaussian approximation in (13) shows how this matrix is governed by the per-subject effect size and by T, rather than by NT. More subjects reduce the sampling variance in (12), but cannot reduce the asymptotic bias originating from uncertainty in the individual classification step. Low-power designs can therefore distort estimated shares regardless of how many subjects the experiment includes.

The proof is in Appendix D.4.

B.2 Mixture estimation

The second procedure estimates shares without assigning a single type to each subject. Instead, each subject receives a probability of belonging to each type, and these latent assignments contribute to the likelihood for the population shares. A common implementation estimates the weights and one within-type parameter for each class jointly:

$$(\hat{w}^{\text{mix}}, \hat{\tilde{\theta}}) \in \arg\max_{w\in\Delta^{K-1},\,\tilde{\theta}_1,\ldots,\tilde{\theta}_K} \sum_{i=1}^{N} \log\left[\sum_{k=1}^{K} w_k \prod_{t=1}^{T} g_k(Y_{it} \mid X_t, \tilde{\theta}_k)\right]. \tag{14}$$

This procedure uses each subject's full record but imposes a common parameter within each type. When the true distributions F_k are point masses, this is the correctly specified finite mixture. However, with parameter heterogeneity, it is generally misspecified. Let $(w^*, \tilde{\theta}^*)$ denote the pseudo-true joint maximizer of the population version of (14). We suppress its dependence on X to simplify notation. For subject i, define the posterior probability of type k' as

$$\tau^*_{ik'} = \frac{w^*_{k'} \prod_t g_{k'}(Y_{it} \mid X_t, \tilde{\theta}^*_{k'})}{\sum_j w^*_j \prod_t g_j(Y_{it} \mid X_t, \tilde{\theta}^*_j)},$$

and the soft confusion matrix

$$Q^*_{kk'} = \mathbb{E}\big[\tau^*_{ik'} \mid Z_i = k\big],$$

where the expectation also averages over $\theta_i \sim F_k$.

For two classes, the connection to design power follows from the posterior log odds. To see this, we define

$$A^*_{ik} \equiv \log \frac{\tau^*_{ik}}{\tau^*_{i,3-k}} = \log \frac{w^*_k}{w^*_{3-k}} + \sum_{t=1}^{T} \log \frac{g_k(Y_{it} \mid X_t, \tilde{\theta}^*_k)}{g_{3-k}(Y_{it} \mid X_t, \tilde{\theta}^*_{3-k})}, \tag{15}$$

then conditional on a type-k subject with parameter θ, let $h(a) = (1 + e^{-a})^{-1}$ and denote the scale and effect size of the posterior log odds by

$$\begin{aligned} \bar{\sigma}^{\mathrm{mix}}_k(\theta; X) &= \sqrt{T^{-1} \operatorname{Var}(A^*_{ik} \mid Z_i = k, \theta_i = \theta)}, \\ \bar{\rho}^{\mathrm{mix}}_k(\theta; X) &= \frac{T^{-1}\mathbb{E}[A^*_{ik} \mid Z_i = k, \theta_i = \theta]}{\bar{\sigma}^{\mathrm{mix}}_k(\theta; X)}. \end{aligned} \tag{16}$$

The numerator has a direct KL interpretation. Under a type-k subject with parameter θ,

$$\begin{aligned} \frac{1}{T}\mathbb{E}[A^*_{ik} \mid Z_i = k, \theta_i = \theta] = \frac{1}{T} \log \frac{w^*_k}{w^*_{3-k}} \\ + \frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}\big(g_k(\cdot \mid X_t, \theta) \,\big\|\, g_{3-k}(\cdot \mid X_t, \tilde{\theta}^*_{3-k})\big) \\ - \frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}\big(g_k(\cdot \mid X_t, \theta) \,\big\|\, g_k(\cdot \mid X_t, \tilde{\theta}^*_k)\big). \end{aligned} \tag{17}$$

The first KL term measures separation from the fitted alternative while the second term subtracts the error from representing a heterogeneous type-k subject with the common parameter $\tilde{\theta}^*_k$. With correctly specified point-mass types, the second term is zero, so this is the original KL separation $\bar{D}_k$ plus the weight-odds term divided by T.

The Gaussian approximation can then be written directly in terms of the effect size,

$$\tau^*_{ik} = h(A^*_{ik}) \approx h\big(\sqrt{T}\, \bar{\sigma}^{\mathrm{mix}}_k(\theta; X)\big[\sqrt{T}\, \bar{\rho}^{\mathrm{mix}}_k(\theta; X) + Z\big]\big), \qquad Z \sim \mathcal{N}(0, 1). \tag{18}$$

Equation (18) makes the role of design power explicit. It uses the same signal-to-noise normalization as $\bar{\rho}_k$ in the main text. With correctly specified point-mass types, it reduces to the original effect size, apart from the mixture-weight log odds divided by T. With parame-

ter heterogeneity, its numerator is the KL distance to the fitted rival minus the KL distance to the fitted version of the subject's own class. Mixture estimation makes soft rather than hard assignments, so the effect size enters through the logistic function. Holding $\bar{\sigma}_k^{\text{mix}}$ fixed, higher $\bar{\rho}_k^{\text{mix}}$ moves posterior weight toward the true type and reduces cross-type confusion. Unlike the hard assignment probability $\Phi(\sqrt{T}\,\bar{\rho}_k^{\text{mix}})$, however, the soft assignment depends on both the effect size and the scale of the log odds.

Proposition 3 (Mixture estimation with common within-type parameters)**.**
Under the usual regularity conditions for maximum likelihood,

$$\hat{w}^{\text{mix}} \xrightarrow{p} w^* = Q^{*\top} w^0, \qquad \sqrt{N}\big(\hat{w}^{\text{mix}} - w^*\big) \xrightarrow{d} \mathcal{N}(0, V_X).$$

Consequently, for each class k, the large-N mean squared error is

$$\text{MSE}(\hat{w}_k^{\text{mix}}) \approx \frac{(V_X)_{kk}}{N} + \big(w_k^* - w_k^0\big)^2.$$

The proof is in Appendix D.5.

If F_k is a point mass at the true parameter for every k, the mixture is correctly specified. The first-order condition for the weights then gives $Q^{*\top} w^0 = w^0$, even though individual posterior probabilities need not be degenerate. Thus, design power affects sampling variance but not the probability limit. If the true F_k is nondegenerate while the estimator imposes a common parameter within each type, the mixture is misspecified and the we generally have $w^* \neq w^0$. Design separation can then affect both asymptotic bias and sampling variance. The intuition for the bias is similar to the previous subsection, which is that under a misspecified model, individual classification is poor when design power is low, and thus bias is introdued to the estimation of aggregate shares. Similarly, this bias cannot be eliminated by increasing the number of subjects.

C Choosing a Design with More Than Two Model Classes

Section 2.2 presents our design criterion for two model classes. When there are $K > 2$ classes, a fixed generating model must be distinguished from $K-1$ rivals, hence its normalized separation $\bar{\rho}_k(\theta_k; \mathbf{X}) = (\bar{\rho}_{kj}(\theta_k; \mathbf{X}))_{j \neq k}$ is a vector with one entry for each rival model class. Since we prefer a scalar design criterion, we first need to turn each $\bar{\rho}_k(\theta_k; \mathbf{X})$ into a scalar by summarizing the pairwise separations.

Our preferred summary is the minimum:

$$\bar{\rho}_k^{\min}(\theta_k; \mathbf{X}) = \min_{j \neq k} \bar{\rho}_{kj}(\theta_k; \mathbf{X}). \tag{19}$$

This is the normalized separation from the rival that is hardest to distinguish from (M_k, θ_k) on design $\mathbf{X}$. The minimum is a natural criterion when the goal is to select M_k over all of its rivals: if one rival remains difficult to distinguish, large separations from the other classes do not resolve that comparison.[25]

The rest of the design recipe follows Section 2.2. For each class k, the researcher specifies a parameter grid $(\theta_{ks})_{s=1}^{S_k}$ with nonnegative weights (w_{ks}) that sum to one, as well as nonnegative model weights $(P_k)_{k=1}^{K}$ that sum to one. The resulting criterion is

$$\mathbf{X}^* \in \arg\max_{\mathbf{X} \in \mathcal{X}_\ell^T} \sum_{k=1}^{K} P_k \sum_{s=1}^{S_k} w_{ks}\, \bar{\rho}_k^{\min}(\theta_{ks}; \mathbf{X}). \tag{20}$$

When $K = 2$, the expression in (20) reduces to (7).

We note that taking the minimum in (19) is a simplification of the vector separation. The minimum does not recover the full power expression in Proposition 1, since it also depends on how the likelihood comparisons against different rivals co-move. The following proposition shows, however, that the minimum bounds approximate power without requiring these correlations.

Proposition 4 (Power bounds with K classes)**.**
Under the conditions of Proposition 1, fix a model (M_k, θ_k) and a design $\mathbf{X}$, and let $n = NT$ and $m = \bar{\rho}_k^{\min}(\theta_k; \mathbf{X})$. For any correlation matrix $\bar{R}_k$,

$$1 - (K-1)\,\Phi\big(-\sqrt{n}\, m\big) \;\le\; \Phi_{\bar{R}_k}\big(\sqrt{n}\, \bar{\rho}_k(\theta_k; \mathbf{X})\big) \;\le\; \Phi\big(\sqrt{n}\, m\big), \tag{21}$$

and the upper bound is attained when the standardized evidence against all rivals is perfectly positively correlated. Consequently, if $\bar{\rho}_k^{\min}(\theta_k; \mathbf{X}_A) > \bar{\rho}_k^{\min}(\theta_k; \mathbf{X}_B)$, then $\mathbf{X}_A$ has higher approximate power than $\mathbf{X}_B$ for all sufficiently large n, regardless of the correlation structures under the two designs.

The proof is in Appendix D.6. To interpret the result, let $q = \Phi(-\sqrt{n}\, m)$ denote the Gaussian approximation to the probability that the hardest-to-distinguish rival beats M_k.

[25] Here the minimum is taken across rival classes for a fixed generating model. This differs from the symmetric objective $\min\{\bar{\rho}_1, \bar{\rho}_2\}$ in Section 4, which takes the minimum across two possible generating models.

Within this approximation, the probability of selecting the wrong class lies between q and $(K-1)q$. Thus $\bar{\rho}_k^{\min}$ determines the rate at which the approximate error probability falls with sample size, up to a factor of at most $K-1$. Correlation across comparisons determines where power lies within these bounds, but it cannot overturn the eventual ranking of two designs whose minimum separations are strictly ordered.

We have used the minimum separation $\bar{\rho}_k^{\min}$, but one could instead work with average separation $\bar{\rho}_k^{\text{avg}} = (K-1)^{-1}\sum_{j\neq k}\bar{\rho}_{kj}$. Because $\bar{\rho}_k^{\text{avg}} \geq \bar{\rho}_k^{\min}$, $\Phi(\sqrt{n}\,\bar{\rho}_k^{\text{avg}})$ is also an upper bound on approximate power, but it is not in general a lower bound. Intuitively, we can have a design where one rival is very challenging to separate, but others are very easy. Modifying the design to increase separation from the "easy-to-distinguish" rivals increases average power; but does not necessarily improve the comparison limiting our probability of correct selection. Since we have formulated our design goal as maximizing the probability of correct selection, we prefer the minimum; a researcher focused on improving a typical pairwise comparison may instead prefer the average. Under the decision-theoretic formulation in Appendix A.1, neither object exactly matches the expected-loss criterion, which uses the multivariate probability of correct selection. Nonetheless, the minimum is a natural scaler proxy which (1) focuses on the hardest comparison and (2) does not require the researcher to specify the correlation structure across all pairwise comparisons.

D Proofs

D.1 Proof of Proposition 1

We give the argument for a fixed design and a fixed scenario (k, θ_k); write $n = NT$. Let q_j^* denote the pseudo-true distribution in class j on the chosen tasks, let θ_j^* be any parameter inducing q_j^*, and let $\hat{\ell}_j$ be the maximized log likelihood per observation. The likelihood regularity condition in Proposition 1(c) is that, jointly over the finitely many model classes,

$$\hat{\ell}_j = \frac{1}{n}\sum_{i=1}^{N}\sum_{t=1}^{T}\log g_j(Y_{it} \mid X_t, \theta_j^*) + o_p(n^{-1/2}), \tag{22}$$

where $\theta_k^* = \theta_k$ for the true class. In words, maximizing over a model's parameters has no first-order effect relative to evaluating its likelihood at a pseudo-true distribution.

For an ordinary finite-dimensional smooth likelihood model, Equation (22) follows from the usual consistency and local quadratic expansion of the likelihood; under standard interiority and nonsingularity conditions, the remainder is in fact $O_p(n^{-1})$; see van der Vaart

(1998, Chapter 5, especially Sections 5.2–5.6). The condition also accommodates a model class that is a finite union of smooth components. A sufficient set of conditions is that all KL-minimizing components induce the same distribution on the chosen tasks, every other component has a strictly positive population-likelihood gap, and the usual likelihood expansion holds within each minimizing component. Uniform convergence then excludes the non-minimizing components with probability approaching one, while Proposition 1(a) ensures that whichever minimizing component is selected has the same first-order likelihood target. In the pure-tremble model in the main text, the preference parameter partitions a fixed design into finitely many choice patterns, so this finite-union argument applies (with the usual smooth likelihood conditions within a pattern if the tremble probability is estimated).

For each subject define the vector of log-likelihood ratios against the projected rivals,

$$\Lambda_i = \left(\frac{1}{T} \sum_{t=1}^{T} \log \frac{g_k(Y_{it} \mid X_t, \theta_k)}{g_j(Y_{it} \mid X_t, \theta_j^*(\theta_k, X))} \right)_{j \neq k}.$$

The vectors Λ_i are i.i.d. across subjects. By Definition 2,

$$\mathbb{E}[\Lambda_i] = \bar{\mathbf{D}}_k, \qquad \mathrm{Var}(\Lambda_i) = \frac{1}{T} \bar{\Sigma}_k, \qquad (\bar{\Sigma}_k)_{jj'} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{Cov}(L_t^{kj}, L_t^{kj'}).$$

The multivariate central limit theorem therefore gives

$$\sqrt{NT} \left(\bar{\Lambda} - \bar{\mathbf{D}}_k \right) \xrightarrow{d} \mathcal{N}(0, \bar{\Sigma}_k), \qquad \bar{\Lambda} = \frac{1}{N} \sum_{i=1}^{N} \Lambda_i. \tag{23}$$

By (22), the vector of fitted likelihood differences satisfies

$$\hat{\mathbf{D}}_k = \bar{\Lambda} + o_p(n^{-1/2}).$$

Combining this expansion with (23) gives

$$\sqrt{NT} \left(\hat{\mathbf{D}}_k - \bar{\mathbf{D}}_k \right) \xrightarrow{d} \mathcal{N}(0, \bar{\Sigma}_k),$$

which proves Proposition 1.

Let $D_k = \mathrm{diag}(\sqrt{\bar{V}_{kj}})_{j \neq k}$. Then $D_k^{-1} \bar{\mathbf{D}}_k = \bar{\rho}_k$ and $D_k^{-1} \bar{\Sigma}_k D_k^{-1} = \bar{R}_k$. Standardizing (23) and using symmetry of a centered Gaussian vector gives the Gaussian power approximation

reported in the main text:

$$\Pr(\hat{k} = k \mid \mathbf{X}, (k, \theta_k)) \approx \Pr\big(W \le \sqrt{NT}\,\bar{\rho}_k(\theta_k; \mathbf{X})\big) = \Phi_{\bar{R}_k}\big(\sqrt{NT}\,\bar{\rho}_k(\theta_k; \mathbf{X})\big),$$

where $W \sim \mathcal{N}(0, \bar{R}_k)$. For $K = 2$, this is the scalar probability $\Phi(\sqrt{NT}\,\bar{\rho}_k)$. If a total-log-likelihood penalty $p_{j,n}$ is subtracted from class j's fitted likelihood, the penalized average gap differs from $\hat{D}_{kj}$ by $(p_{k,n} - p_{j,n})/n$. Hence differences in penalties that are $o(\sqrt{n})$ are $o(n^{-1/2})$ on the average-likelihood scale and do not change the first-order distribution or the resulting power approximation.

D.2 Proof of Lemma 1

Write $r_A = \bar{\rho}(X_A)$, $r_B = \bar{\rho}(X_B)$, and $g = 1 - r_B/r_A$. For $x \ge 0$, the standard normal CDF satisfies

$$\Phi(x) - \tfrac{1}{2} \le \phi(0)x, \qquad \Phi(x) - \tfrac{1}{2} \ge \phi(0)x\left(1 - \frac{x^2}{6}\right).$$

The second inequality follows by integrating the bound $e^{-u^2/2} \ge 1 - u^2/2$; the first follows from concavity of Φ on $[0, \infty)$. By assumption, every scenario-specific effect size under X_A satisfies $\sqrt{n}\,\bar{\rho}_s(X_A) < z_\beta$. The gap condition implies $z_\beta^2 \le 6g$, so the lower bound applied scenario by scenario to X_A and the upper bound applied to X_B yield

$$\mathrm{WAP}(X_A; n) - \tfrac{1}{2} \ge \phi(0)\sqrt{n}\,r_A(1 - g) = \phi(0)\sqrt{n}\,r_B \ge \mathrm{WAP}(X_B; n) - \tfrac{1}{2}.$$

The first inequality is strict because $r_A > 0$ and the separation condition is strict. Hence $\mathrm{WAP}(X_A; n) > \mathrm{WAP}(X_B; n)$.

D.3 Proof of Corollary 1

Consider first truth $(1, \theta_1)$. The proof of Proposition 1 shows, for two models, that

$$\hat{D} = \bar{D}_{12}(\theta_1; X) + O_p(n^{-1/2}).$$

Finite second moments and a positive separated variance imply that the sample variance used in $\hat{\sigma}^2$ converges to $\bar{V}_{12}(\theta_1; X)$. Therefore

$$\hat{z} = \sqrt{n}\,\bar{\rho}_{12}(\theta_1; X) + O_p(1).$$

Under truth $(2, \theta_2)$, reversing the likelihood ratio changes both centers' signs, giving the first part of the corollary.

For the second part, integrate these conditional limits over the model and scenario weights. The assumed moment bounds give uniform integrability, so the limits can be taken through expectations. Under truth 1, the signed mean is $\mathbb{D}_1$ and the absolute mean is also $\mathbb{D}_1$; under truth 2, the signed mean is $-\mathbb{D}_2$ and the absolute mean is $\mathbb{D}_2$. The same argument applied to $\hat{z}/\sqrt{n}$ gives $\bar{\rho}_1$ and $-\bar{\rho}_2$. Weighting the two cases by P_1 and P_2 gives the four limits in Corollary 1.

D.4 Proof of Proposition 2

Holding T fixed, the assignment indicators are i.i.d. across subjects, with $\Pr(\hat{Z}_i = k') = (Q^\top w^0)_{k'}$. The law of large numbers therefore gives $\hat{w}^{\text{cls}} \to_p Q^\top w^0$. For each class k, the indicators $\mathbf{1}\{\hat{Z}_i = k\}$ are Bernoulli with success probability $r_k = (Q^\top w^0)_k$, so

$$\text{MSE}(\hat{w}_k^{\text{cls}}) = \frac{r_k(1-r_k)}{N} + (r_k - w_k^0)^2.$$

For the second result, fix a type k and parameter θ. Under the stated likelihood regularity conditions, maximizing each within-subject likelihood has no first-order effect relative to evaluating it at its pseudo-true distribution, so

$$\hat{D}_{ik}^{\text{cls}} = \frac{1}{T}\sum_{t=1}^{T} L_t^{k,3-k} + o_p(T^{-1/2}).$$

The task-level log-likelihood ratios have total mean and variance, respectively,

$$T\bar{D}_{k,3-k}(\theta; X) \qquad \text{and} \qquad T\bar{V}_{k,3-k}(\theta; X).$$

The Lindeberg central limit theorem and the preceding likelihood expansion therefore give

$$\frac{\sqrt{T}\left(\hat{D}_{ik}^{\text{cls}} - \bar{D}_{k,3-k}(\theta; X)\right)}{\sqrt{\bar{V}_{k,3-k}(\theta; X)}} \xrightarrow{d} \mathcal{N}(0,1).$$

Because the classifier assigns the subject to class k when $\hat{D}_{ik}^{\text{cls}} \geq 0$, the corresponding Gaussian approximation is

$$\Pr(\hat{Z}_i = k \mid Z_i = k, \theta_i = \theta) \approx \Phi\big(\sqrt{T}\,\bar{\rho}_{k,3-k}(\theta; X)\big).$$

Averaging over $\theta \sim F_k$ gives (13); for two classes, the off-diagonal probability is its complement.

D.5 Proof of Proposition 3

Let $\psi = (w, \tilde{\theta})$ collect the weights and common within-type parameters. The joint mixture estimator is an ordinary, possibly misspecified, M-estimator based on independent subject records. The law of large numbers gives $\hat{\psi} \to_p \psi^* = (w^*, \tilde{\theta}^*)$. The usual score expansion and subject-level central limit theorem give asymptotic normality; V_X is the weight block of the resulting sandwich covariance matrix. For the weights, the population score (using a Lagrange multiplier for the simplex constraint) is equivalent to

$$w^*_{k'} = \mathbb{E}\big[\tau^*_{ik'}\big] = \sum_{k=1}^{K} w^0_k Q^*_{kk'},$$

which gives $w^* = Q^{*\top} w^0$. If the mixture is correctly specified, the posterior probabilities satisfy the usual iterated-expectations identity, so the right-hand side equals w^0. Finally,

$$\mathbb{E}\big[(\hat{w}^{\text{mix}}_k - w^0_k)^2\big] = \text{Var}(\hat{w}^{\text{mix}}_k) + \big(\mathbb{E}[\hat{w}^{\text{mix}}_k] - w^0_k\big)^2 \approx \frac{(V_X)_{kk}}{N} + (w^*_k - w^0_k)^2,$$

which is the MSE expression in Proposition 3.

D.6 Proof of Proposition 4

Let $W \sim \mathcal{N}(0, \bar{R}_k)$ and $v = \sqrt{n}\,\bar{\rho}_k(\theta_k; \mathbf{X})$, so that $\Phi_{\bar{R}_k}(v) = \Pr(W_j \le v_j \text{ for all } j \ne k)$ and each W_j is standard normal. The event $\{W_j \le v_j \text{ for all } j \ne k\}$ is contained in $\{W_{j'} \le v_{j'}\}$ for the rival j' attaining $\min_j v_j = \sqrt{n}\,m$, which gives the upper bound. The union bound gives

$$\Pr(W_j \le v_j \text{ for all } j \ne k) \ge 1 - \sum_{j \ne k} \Phi(-v_j) \ge 1 - (K-1)\,\Phi(-\sqrt{n}\,m),$$

which is the lower bound. If every entry of $\bar{R}_k$ equals one, all W_j coincide almost surely, so $\Pr(W_j \le v_j \text{ for all } j \ne k) = \Phi(\min_j v_j) = \Phi(\sqrt{n}\,m)$ and the upper bound is attained. This is the case when the log-likelihood ratios L_t^{kj} coincide across j, for instance when every rival's pseudo-true model induces the same distribution on the tasks in $\mathbf{X}$.

For the last claim, write m_A and m_B for the two minima, with $m_A > m_B \ge 0$. Applying the lower bound to $\mathbf{X}_A$ and the upper bound to $\mathbf{X}_B$, the difference in approximate

power is at least $\Phi(-\sqrt{n}\,m_B)-(K-1)\,\Phi(-\sqrt{n}\,m_A)$, which is positive for all large n because $\Phi(-\sqrt{n}\,m_A)/\Phi(-\sqrt{n}\,m_B)\to 0$.

E Algorithmic Design

Let $f : \mathcal{X}_\ell^T \to \mathbb{R}$ denote the objective in Equation (7), taken as a deterministic function of the design space $\mathbf{X}$. Our implementation follows the TuRBO algorithm of Eriksson et al. (2019), up to the minor implementation changes. To begin, we rescale the design space coordinates to the unit hypercube $[0,1]^d$ and maintain a single adaptive trust region.

The main distinction between TuRBO and conventional Bayesian Optimization is that local trust regions are maintained in the optimization process to avoid difficulties caused by global idiosyncrasies of the objective function. Within the trust region, we fit a Gaussian-process (GP) model using only evaluations made since the region was initialized. In a GP, f is treated as an unknown function which is modeled probabilistically. When the GP learns the value of f at specific points (designs), it uses this information to form a predictive distribution of f at nearby points which have not been evaluated. We use a Matérn-5/2 covariance function and separate length scales for each design dimension. We also standardize f prior to fitting the GP. Given observations $\mathbf{y}=(y_1,\ldots,y_t)^\top$, let K be their covariance matrix and let $\mathbf{k}(\mathbf{X})$ contain their covariances with a new design $\mathbf{X}$. The standard GP posterior has mean and covariance

$$\mu_t(\mathbf{X}) = m + \mathbf{k}(\mathbf{X})^\top (K+\lambda I)^{-1}(\mathbf{y}-m\mathbf{1}),$$
$$k_t^{\text{post}}(\mathbf{X},\mathbf{X}') = k(\mathbf{X},\mathbf{X}') - \mathbf{k}(\mathbf{X})^\top (K+\lambda I)^{-1}\mathbf{k}(\mathbf{X}'),$$

where m is a fitted constant and λ is a small noise term used to stabilize the fit. The posterior therefore provides both a predicted objective value and uncertainty about that prediction.

The trust region is a hyperrectangle centered at the best design within it. If L is its base side length and ℓ_j is the GP length scale for coordinate j, its side length in that coordinate, before intersecting the region with $[0,1]^d$, is $L_j = L\ell_j/(\prod_{h=1}^d \ell_h)^{1/d}$. Within this region, we generate candidate designs by changing each coordinate of the center with probability $\min\{1, 20/d\}$, while requiring every candidate to change at least one coordinate. To select a batch, we draw candidate values jointly from the GP posterior and select the candidate with the largest value in each draw, without selecting a candidate twice. Again, this form of Thompson sampling favors designs that either have high predicted values or remain sufficiently uncertain, striking a balance between exploration and exploitation.

The base side length begins at $L = 0.8$. It is doubled after consecutive successful batches and halved after consecutive unsuccessful batches, where success means improving the region's best value by more than 10^{-3} times its absolute value. If the region persistently see no success and is shrunk to $L < 2^{-7}$, then it is terminated, and we draw a new Latin-hypercube initialization over the full design space. The GP then uses only data from the new region, while the best design found across all regions is retained.[26]

To assess the performance of our preferred algorithm, we conduct a computational benchmark on two objectives: one-sided CPT truth against a pseudo-true EU alternative, and random-preference CPT (RP-CPT) truth against a pseudo-true random-preference EU (RP-EU) alternative. We consider designs with $T \in \{2, 4, 8, 12\}$ two-state lottery-choice tasks. Each task contributes five continuous design coordinates, so the search dimension ranges from 10 to 60. The objective averages the one-sided normalized separation $\bar{\rho}$ over six parameter values from the maintained prior for the true model.

We compare the performance of four methods after the same initializing grid evaluation, including (i) our reference-style TuRBO-1 implementation; (ii) a single BFGS search in logit-transformed coordinates[27], initialized at the best design from the common initialization; (iii) Nelder–Mead in logit-transformed coordinates, using ten starts selected from the best designs in the common initialization; and (iv) randomized maximin Latin-hypercube (LHS) search. [28]

All benchmark methods search over continuous, common-state-space lotteries. Probabilities lie in $[0.05, 0.95]$ and payoffs lie in $[0, 100]$. Within each task, we parameterize payoffs to make lottery A weakly better in one state and lottery B weakly better in the other, thereby excluding strict dominance. Within each model-pair, task-count, and seed cell, all methods receive the same deterministic 100-point LHS initialization. This initialization is included in a maximum total budget of 1,600 objective evaluations. TuRBO uses batches of $q = 8$ and $C = 500$ candidates per batch. The inner pseudo-true-parameter calculation uses five baseline starting values together with deterministic boundary starts, for a total of seven starts for EU and nine for RP-EU.[29] We use the finalized parameter bounds except for CPT versus EU at $T = 2$, where the EU search is widened to $\alpha \geq 0.05$ and $\eta \leq 50$. When a

[26]Our implementation differs from the original Eriksson et al. (2019) in that we use a randomized Latin hypercube rather than a scrambled Sobol sequence to generate candidates, and we use a smaller number of $C = 500$ candidates. We use `GauPro` to fit the GP.

[27]We use logit-transformation so that BFGS can search over an unconstrained space instead of bounded within $[0, 1]$.

[28]For our design space, a literal grid search is practically infeasible since the dimensionality is high, hence we use maximin LHS as the grid-like benchmark.

[29]This is to avoid the inner pseudo-true-parameter optimizer getting stuck in local solutions.

local optimizer terminates before exhausting its allocation, we start a new round and count toward the 1,600-evaluation horizon.

Figure 9 summarizes the results. TuRBO-1 has the highest mean final objective among the displayed methods in every $T \geq 4$ comparison, with its advantage becoming especially pronounced as the parameter dimension increases. For smaller designs at $T = 2$, all four methods essentially approach the same ceiling. Therefore our computation suggests that TuRBO-1 performs well for the higher-dimensional problems that motivate our algorithm.

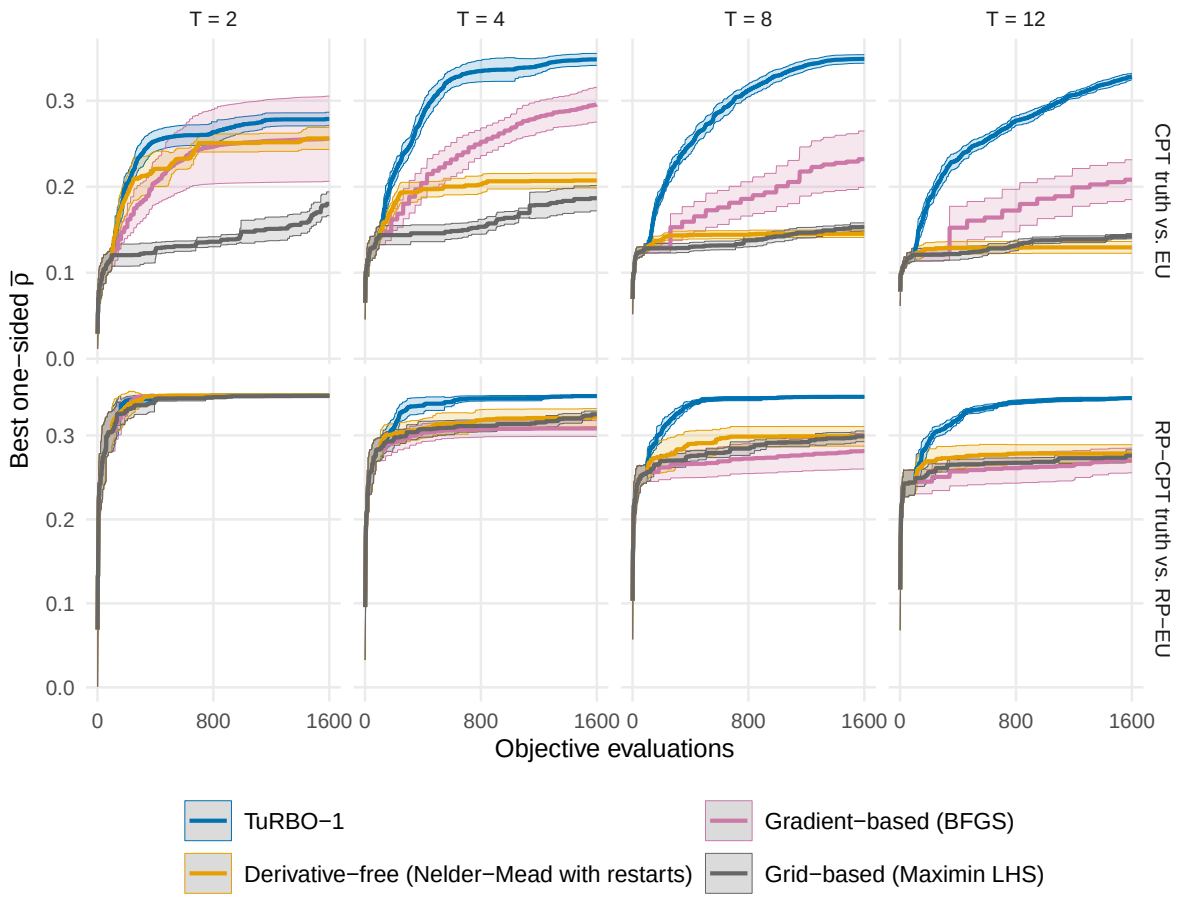


Figure 9: Optimizer convergence on the paper's one-sided normalized-separation objective. Curves show the mean incumbent and shading gives pointwise 95% t confidence intervals for the mean across twelve seeds. The four displayed methods represent Bayesian optimization, gradient-based, derivative-free, and space-filling search.

F Experimental Design

F.1 Parametric Theories and Priors

We use six different parametric theories of risk preference to guide our experimental design. These functional forms, as well as the prior parameter values used to calculate the design criterion, are preregistered at https://osf.io/xyab3/overview. We reproduce both below. To get from the parameter priors to our design criterion, we put equal weights on the given parameter combinations.

Each model specifies binary choice probabilities between two lotteries, which we label A and B. A generic lottery L is described by a payoff vector x_L and a corresponding probability vector p_L. We use x_L^i to indicate the lottery's payout in the i'th state, which occurs with probability p_L^i. We use $s(L)$ to denote the number of states in lottery L. For pairs of lotteries A and B in our choice problems, a common state-space is explicitly specified, i.e. $s(A) = s(B)$ and $p_A = p_B$. Additionally, all payoffs are positive in our dataset.

Expected Utility Theory (EUT)

$$\Pr(A) = \frac{1}{1 + \exp(\eta \cdot (V(B) - V(A)))},$$

where

$$V(L) = \sum_{i=1}^{s(L)} (x_L^i)^\alpha \cdot p_L^i$$

Since we never place any weight on EUT when calculating our design criterion (instead we place full weight on the nesting model CPT), we do not need to specify priors on parameters.

Cumulative Prospect Theory (CPT)

CPT builds on EUT by introducing a Prelec probability weighting function.

For a lottery L with $s(L)$ states, let $\pi_L : \{1, \ldots, s(L)\} \to \{1, \ldots, s(L)\}$ be the permutation that ranks outcomes in ascending order, so $x_L^{\pi_L(1)} \leq x_L^{\pi_L(2)} \leq \cdots \leq x_L^{\pi_L(s(L))}$. Using the one-parameter Prelec (1998) probability weighting function

$$w(p; \gamma) = \exp(-(-\ln p)^\gamma), \qquad \gamma > 0,$$

the decision weight on the k-th ranked outcome of lottery L is the difference of weighted

decumulative probabilities,

$$v_L^k(\gamma) = w\left(\sum_{j=k}^{s(L)} p_L^{\pi_L(j)}; \gamma\right) - w\left(\sum_{j=k+1}^{s(L)} p_L^{\pi_L(j)}; \gamma\right),$$

with the convention that the second sum is zero when $k = s(L)$.

As with EUT, we have

$$\Pr(A) = \frac{1}{1 + \exp(\eta \cdot (V(B) - V(A)))},$$

where now

$$V(L) = \sum_{k=1}^{s(L)} v_L^k(\gamma) \cdot \left(x_L^{\pi_L(k)}\right)^\alpha.$$

Our priors cover the parameter combinations

$$\begin{aligned}(\alpha, \gamma, \eta) \in \{&(0.6, 0.4, 0.58),\\ &(0.8, 0.4, 0.23),\\ &(0.6, 0.6, 0.73),\\ &(0.8, 0.6, 0.33),\\ &(0.6, 0.8, 0.76),\\ &(0.8, 0.8, 0.38)\}\end{aligned}$$

Random Preference Expected Utility (REU)

For each individual decision, a power utility curvature parameter α is drawn from a lognormal distribution with median m_α and log-scale σ_α,

$$\ln \alpha \sim N(\ln m_\alpha, \sigma_\alpha^2), \qquad \alpha > 0,$$

where m_α is the median curvature.[30]

Given α, the lotteries are evaluated according to expected utility,

$$V(L \mid \alpha) = \sum_{j=1}^{s(L)} (x_L^j)^\alpha \cdot p_L^j,$$

[30]To speed computation, we truncate the distribution of α at 0.01 on the low end and 10 at the high end (with appropriate renormalization).

and A is chosen whenever $V(A \mid \alpha) > V(B \mid \alpha)$.

To allow for dominance violations and other noise, we add a tremble: with probability $\omega \in (0,1)$ the choice is made uniformly at random, and with probability $1-\omega$ the subject follows their expected utility ranking. Letting $\Delta(\alpha) \equiv V(A \mid \alpha) - V(B \mid \alpha)$, the choice probability conditional on α is

$$\Pr(A \mid \alpha) = (1-\omega)\,\mathbf{1}\{\Delta(\alpha) > 0\} + \tfrac{\omega}{2}.$$

Integrating over the distribution of α, the unconditional choice probability for a given problem is

$$\Pr(A) = (1-\omega)\Pr_{\alpha \sim \text{LogN}(\ln m_\alpha, \sigma_\alpha^2)}[\Delta(\alpha) > 0] + \tfrac{\omega}{2}.$$

The estimable parameters are $(m_\alpha, \sigma_\alpha, \omega)$.

Again, we never place weight on this theory for our design criterion (all weight is on the nesting theory RCPT), so we do not specify priors.

Random Preference CPT (RCPT)

In random preference CPT (RCPT), we modify random expected utility by introducing a single (non-random) probability weighting parameter.

Given α, each lottery is evaluated according to its rank-dependent CPT value,

$$V(L \mid \alpha, \gamma) = \sum_{k=1}^{s(L)} \nu_L^k(\gamma) \cdot \left(x_L^{\pi_L(k)}\right)^\alpha,$$

and A is chosen whenever $V(A \mid \alpha, \gamma) > V(B \mid \alpha, \gamma)$.

As before, with probability $\omega \in (0,1)$ the choice is made uniformly at random, and with probability $1-\omega$ the choice matches the CPT ranking given α. Letting $\Delta(\alpha;\gamma) \equiv V(A \mid \alpha, \gamma) - V(B \mid \alpha, \gamma)$, the unconditional choice probability is

$$\Pr(A) = (1-\omega)\Pr_{\alpha \sim \text{LogN}(\ln m_\alpha, \sigma_\alpha^2)}[\Delta(\alpha;\gamma) > 0] + \tfrac{\omega}{2}.$$

Our priors cover the parameter combinations

$$\begin{aligned}(m_\alpha, \sigma_\alpha, \gamma, \omega) \in \{&(0.6, 0.4, 0.6, 0.4),\\ &(0.8, 0.4, 0.6, 0.4),\\ &(0.6, 0.6, 0.6, 0.4),\\ &(0.8, 0.6, 0.6, 0.4),\\ &(0.6, 0.8, 0.6, 0.4),\\ &(0.8, 0.8, 0.6, 0.4)\}\end{aligned}$$

Comparison Complexity (CC)

Under the comparison complexity model from Shubatt and Yang (2023), lotteries are evaluated according to expected utility, and binary choice noise is an increasing function of the CDF dissimilarity between the lotteries.

As in EUT, we have

$$V(L) = \sum_{i=1}^{s(L)} u(x_L^i) \cdot p_L^i,$$

where $u(x) = x^\alpha$.

Let F_L denote the CDF of lottery L over outcomes and F_L^{-1} its quantile function. The CDF-distance between lotteries A and B is

$$d^{CDF}(A, B) = \int_0^1 \left| u(F_A^{-1}(q)) - u(F_B^{-1}(q)) \right| dq,$$

which measures the area between the utility-valued quantile functions of A and B. The value-dissimilarity ratio is

$$R^{CDF}(A, B) = \frac{V(A) - V(B)}{d^{CDF}(A, B)},$$

which takes values in $[-1, 1]$ by construction, with $R^{CDF} = 1$ (-1) when A (B) first-order stochastically dominates the other lottery.

Following Shubatt and Yang (2026), the choice probability is a monotone transformation of R^{CDF},

$$\Pr(A) = G\left(R^{CDF}(A, B; \alpha)\right),$$

with

$$G(r) = \begin{cases} (1-\kappa) - (0.5-\kappa)(1-r)^{\gamma} & \text{if } r \geq 0, \\ \kappa + (0.5-\kappa)(1+r)^{\gamma} & \text{if } r < 0, \end{cases}$$

where $\kappa \in [0, 1/2)$ is a tremble parameter governing error rates at dominance and $\gamma > 0$ governs the curvature of the response function in the ratio.

The estimable parameters are (α, γ, κ).

Our priors cover the parameter combinations

$$\begin{aligned} (\alpha, \kappa, \gamma) \in \{&(0.6, 0.03, 0.5), \\ &(0.8, 0.03, 0.5), \\ &(0.6, 0.1, 0.5), \\ &(0.8, 0.1, 0.5), \\ &(0.6, 0.03, 0.85), \\ &(0.8, 0.03, 0.85), \\ &(0.6, 0.1, 0.85), \\ &(0.8, 0.1, 0.85) \end{aligned}$$

Salience (SAL)

We use the continuous version of the salience model of choice under risk from Bordalo et al. (2013). When a lottery is evaluated according to salience, the decision weight placed on each state is distorted according to how *salient* that state is.

The salience value for a lottery A in the menu $\{A, B\}$ is

$$V(A; B) = \sum_{i=1}^{s(A)} \omega_i(A, B) \cdot p_i \cdot x_A^i,$$

where $\omega_i(A, B)$ is a salience distortion term given by

$$\omega_i(A, B) = \frac{\delta^{-\sigma(x_A^i, x_B^i)}}{\sum_{j=1}^{s(A)} \delta^{-\sigma(x_A^j, x_B^j)} p_j},$$

and σ is the symmetric salience function[31]

$$\sigma(y,z) = \frac{|y-z|}{|y|+|z|+1}$$

Assuming logit noise,

$$\Pr(A) = \frac{1}{1+\exp(\eta \cdot (V(B;A) - V(A;B)))}.$$

Our priors cover the parameter combinations

$$(\delta,\eta) \in \{(0.01, 0.042),$$
$$(0.1, 0.086)$$

[32]

F.2 Experimental Instructions

[31] In Bordalo et al. (2012), the salience function takes a parameter θ; we have restricted this value to 1, due to weak identification. In pilot data, allowing θ to be estimated offered only very minor model fit improvements.

[32] We roughly calibrated our prior on δ using the common consequence problem studied in Section V of Bordalo et al. (2012).

Instructions I: Overview

Please read these instructions carefully. There will be comprehension checks. If you fail these checks, you will be excluded from the study and you will not receive the completion payment.

In this study, you will make multiple decisions.

Your payment will consist of two components:

- Completion fee:

  If you pass all our **comprehension checks and complete the study**, you will receive a **completion fee of $4.50.**

- Additional bonus:

  In each of 50 rounds, you will make a decision. One of the decisions will be selected at random by the computer, with equal probability, and will determine your **bonus**. The **maximal bonus** you can earn in this study is **$50**. Your bonus will actually be paid out to you with probability 0.1.

Instructions II: Lottery Pairs

Throughout this study, you will encounter pairs of lotteries. The lotteries, called **Option A** and **Option B,** will each pay out money depending on the color of a randomly drawn ball. We will present these lotteries to you in a table like the one below:

Probability		Option A	Option B
	40%	$12	$24
	50%	$60	$89
	10%	$68	$11

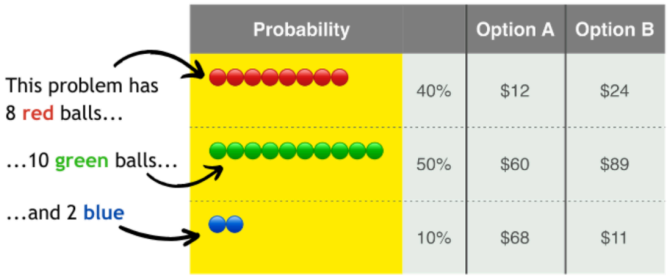
Instructions II: Lottery Pairs

The first two columns tell us the *likelihood* of drawing each colored ball. For each lottery pair, there are 20 balls total, one of which will be drawn at random to determine payouts.

Probability		Option A	Option B
	40%	$12	$24
	50%	$60	$89
	10%	$68	$11

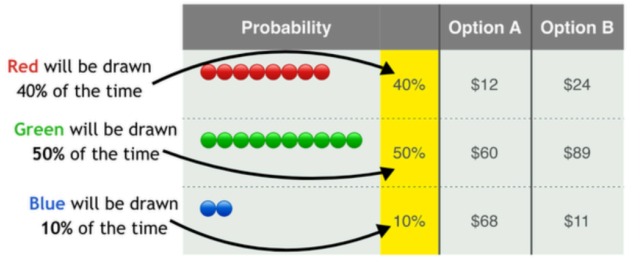
Instructions II: Lottery Pairs

The first column depicts the number of red, green, and blue balls.



Instructions II: Lottery Pairs

The second column expresses the same information as probabilities.



Instructions II: Lottery Pairs

The third and fourth columns tell us how much the options pay out for each ball color.



Instructions III: Study Tasks

Task 1

For each pair of lotteries you encounter, you will be asked to **click on the option you would prefer to receive.** For example, if you liked **Option B** better than **Option A**, you would click the button for **Option B.**



There are no right or wrong answers. We are just interested to know which you option you would prefer.

Instructions IV: Attention Checks

Throughout the study, you will encounter attention checks. Be sure to read each page carefully in order to pass these checks. If you fail an attention check, you will not be eligible for a bonus payment.

Towards the end of the survey, you will be specifically asked to enter the **key phrase** from the instructions. The key phrase is **"roses are red".** We recommend noting this down so that you can access it it later.

Instructions V: Bonus Payment

If you complete the entire study and pass all attention checks, you will have the chance to earn a bonus.

If you are selected for a bonus, the computer will randomly select **one of the 50 choices** you saw and simulate drawing a red, green, or blue ball, according to the stated probabilities. If a red ball is drawn, you will actually receive the "red" payment for the option you chose; if a green ball is drawn, you will receive the "green" payment, and so on.

On average, **these bonus payments are worth $25**. If you successfully complete the study, you'll have a **1 in 10 chance** of being selected for a bonus.

F.3 Aggregated Experimental Specifications

In the main text, we separately evaluated the relationship between power and strength and power and direction for each theory pair, on all 60 experimental designs. This is the most direct test implied by our theory, and by evaluating each pair separately, we can clearly see where the relationships are strongest. We additionally pre-registered an aggregate analysis where we assigned each design to a particular theory-pair and assessed the relationship between $\bar{\rho}$ and the log likelihood ratio. Here, we describe exactly how we conducted these analyses and report results. The reported correlations are significant at $p = 0.05$ and qualitatively consistent with Table 3 in the main text.

To test whether more powerful designs lead to stronger statistical evidence, we can compare the three optimized designs from the *one-sided* and *symmetric* objectives and see how results compare to alternative designs which were *not* optimized for power. We could achieve this by constructing and running minimally-powered designs for each objective; but this seemed to us an unfair comparison. Instead, for each theory pair, we selected three designs from the set of 36 *asymmetric* designs such that the two theories were similarly powered (for nested pairs, we didn't need this condition). So for instance, an *asymmetric* design which is highly powered to separate Comparison Complexity from Random Cumulative Prospect Theory turns out to be a symmetric but low-powered design to separate between Comparison Complexity and Cumulative Prospect Theory; thus, we assign it as a CC-CPT design for this first empirical exercise. We prefer this to minimally-powered designs because these asymmetric designs are not obviously useless – in fact they are explicitly very useful for some alternative theory tests. But, they are by definition *not* the best designs for testing the theory pair at hand.

After applying this assignment scheme, we end up with a set of 6 designs for each objective: 3 algorithmically optimized for high power, and 3 reasonable designs constructed for another purpose. Our pre-registration specifies this design-to-theory assignment scheme, and reports calculated $\bar{\rho}_1$ and $\bar{\rho}_2$ for each design. We then conduct a simple test: in this sample of 48 designs, does the absolute value of the LLR (i.e., the size of the winning model's advantage) increase with design power? For symmetric theory pairs, we define design power as the average of $\bar{\rho}_1$ and $\bar{\rho}_2$; for one-sided pairs where M_1 is the nesting model, we just use $\bar{\rho}_1$. Figure 10 presents the results visually. The correlation is indeed positive ($r = 0.41$) and statistically significant ($p = 0.004$).

To test how asymmetry in design power affects empirical conclusions, we can compare our Asymmetric 1 and Asymmetric 2 designs from Table 2. For six different theory pairs,

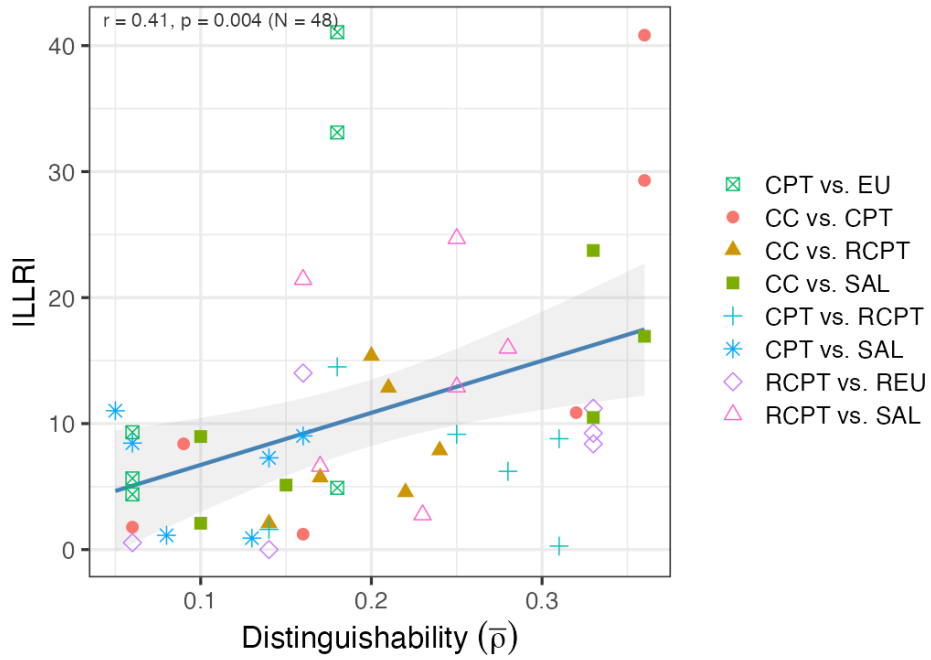


Figure 10: Strength of statistical evidence as a function of design power

we ran three designs with large $\bar{\rho}_1$ and small $\bar{\rho}_2$ (Asymmetric 1); and three with small $\bar{\rho}_1$ and large $\bar{\rho}_2$ (Asymmetric 2). We take these 36 designs and regress the log likelihood ratio on $\Delta\bar{\rho} = \bar{\rho}_1 - \bar{\rho}_2$. Figure 11 presents the results. There is a striking positive correlation ($r = 0.81, p < 0.001$). For every single theory pair, evidence indeed looks more favorable for Theory 1 on the designs asymmetrically powered for Theory 1 ($\Delta\bar{\rho} > 0$); and more favorable for Theory 2 on the designs asymmetrically powered for Theory 2 ($\Delta\bar{\rho} < 0$). For some pairs, such as CC vs. RCPT, we see exactly the scenario we described above: Comparison Complexity beats Random CPT on the designs powered only for Comparison Complexity; but Random CPT beats Comparison Complexity on the designs powered only for Random CPT.

F.4 Robustness to Vuong Degeneracy

The classical Vuong statistic used in Tables 3 and 4 can sometimes be degenerate when fitted model classes overlap and are equally close to the realized empirical distribution. This changes the relevant distribution of the statistic and hence affects interpretation of magnitudes. In particular, comparing magnitudes of the statistics between degenerate and non-degenerate typically does not give rise to any clean interpretation. To protect against such issues we repeat the specifications in Tables 3 and 4 that use the Vuong statistic as the

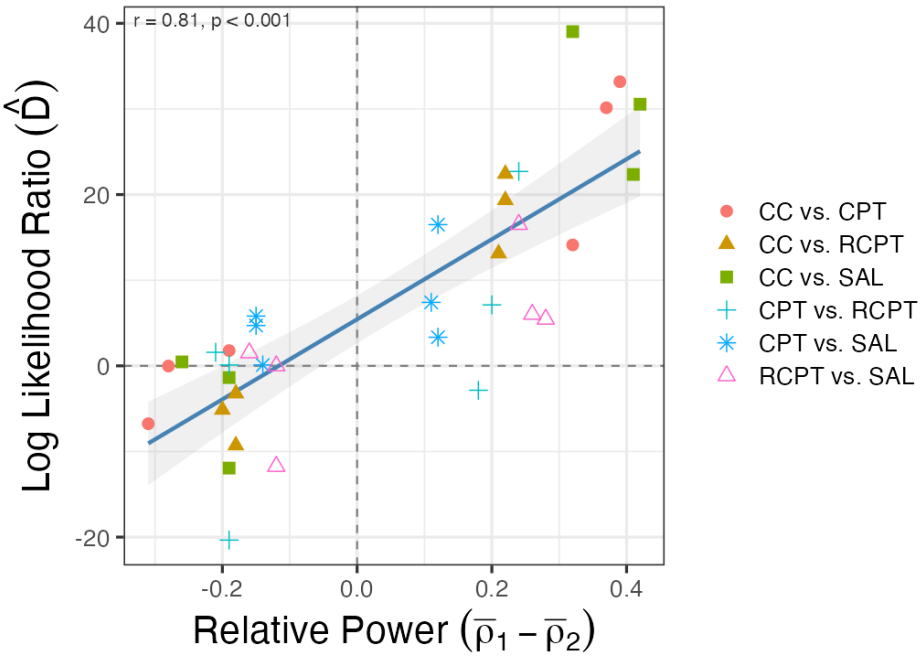


Figure 11: Direction of statistical evidence as a function of relative design power

empirical outcome and replace it with the one-step non-degenerate statistic of Shi (2015). This statistic is modified based on the Vuong statistic and designed to control size uniformly in both overlapping and non-overlapping cases. For each model pair, we regress $|\hat{t}_{\mathrm{ndV}}|$ on $\sqrt{NT}\,\bar{\rho}$; for the six non-nested pairs, we also regress $\hat{t}_{\mathrm{ndV}}$ on $\sqrt{NT}\,\bar{\rho}_{\Delta}$. All 480 design–model-pair cells yield a finite statistic. The design sample, regressors, and transformations otherwise match Tables 3 and 4.

All estimated coefficients remain positive. Five of the eight strength coefficients and four of the six directional coefficients are statistically significant at the five-percent level. Thus, the empirical relationships between ex ante normalized separation and ex post model-selection evidence do not appear to be driven by the classical Vuong normalization.

Theory pair	$\lvert\hat{t}_{\text{ndV}}\rvert \sim \sqrt{NT}\,\bar{\rho}$	$\hat{t}_{\text{ndV}} \sim \sqrt{NT}\,\bar{\rho}_{\Delta}$
CPT vs. EU	0.69***	
	(0.20)	
RCPT vs. REU	0.37***	
	(0.07)	
CC vs. CPT	0.48***	1.07***
	(0.09)	(0.16)
CC vs. RCPT	0.12	0.62***
	(0.15)	(0.18)
CC vs. SAL	0.63***	1.14***
	(0.11)	(0.12)
CPT vs. RCPT	0.10	0.67*
	(0.15)	(0.36)
CPT vs. SAL	0.55**	0.51
	(0.24)	(0.33)
RCPT vs. SAL	0.13	0.83**
	(0.20)	(0.40)

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$

Table 5: Regressions corresponding to column (1) of Tables 3 and 4, using the non-degenerate Vuong statistic of Shi (2015) as the empirical outcome. Each regression contains all 60 designs for the indicated model pair. Parentheses report OLS standard errors; significance stars refer to tests that the regression slope equals zero. Directional regressions are not defined for the two nested pairs under our one-sided design objective.