

How Flexible is that Functional Form?

Quantifying the Restrictiveness of Theories*

Drew Fudenberg[†] Wayne Gao[‡] Annie Liang[§]

August 15, 2021

Abstract

We propose a new way to quantify the restrictiveness of an economic model, based on how well the model fits simulated, hypothetical data sets. The data sets are drawn at random from a distribution that satisfies some application-dependent content restrictions, such as that people prefer more money to less. Models that can fit almost all hypothetical data well are not restrictive. To illustrate our approach, we evaluate the restrictiveness of popular models in two laboratory settings—certainty equivalents and initial play—and in one field setting—takeup of microfinance in Indian villages. We explain how restrictiveness reveals new insights about each of the models.

Keywords: model flexibility, parsimony, synthetic data, behavioral and experimental economics

*We thank Nikhil Agarwal, Victor Aguiar, Abhijit Banerjee, Tilman Börgers, Vince Crawford, Glenn Ellison, Ben Golub, Taisuke Imai, Shaowei Ke, David Laibson, Rosa Matzkin, John Quah, Karen Rozen, Charles Sprenger, Dmitry Taubinsky, and Emanuel Vespa for helpful comments, and NSF grants SES 185162 and 1951056 for financial support.

[†]Department of Economics, MIT; drew.fudenberg@gmail.com

[‡]Department of Economics, University of Pennsylvania; waynegao@upenn.edu

[§]Department of Economics and Department of Computer Science, Northwestern University; an-nie.liang@northwestern.edu.

1 Introduction

If a parametric model fits the available data well, is it because the model captures structure that is specific to the observed data, or because the model is so flexible that it would fit almost all conceivable data? This paper provides a quantitative measure of model restrictiveness that can distinguish between these two explanations, and is easy to compute across a variety of applications.

Our approach for evaluating the restrictiveness of a given model is as follows: The analyst first imposes some background constraints on the data that reflect prior knowledge or intuition about the setting, such as that people prefer more money to less. If it is understood that the model satisfies these basic properties, then the question of interest is how much additional content it contains. To get a measure of this, we propose generating random “admissible” data sets that satisfy the specified constraints. Our measure of the model’s restrictiveness is its normalized average error across the admissible data sets. An unrestrictive model fits most of the admissible data and thus does not add much to the background constraints. In contrast, a restrictive model performs poorly on the admissible data sets, so in this case the model does impose additional constraints.

While restrictive models are desirable, a restrictive model is not particularly useful if it rules out the regularities that are present in real data. We thus complement the evaluation of restrictiveness, which is based solely on synthetic data, with an evaluation of the model’s performance on actual data, using the measure of completeness proposed in Fudenberg et al. (2021). Together, restrictiveness and completeness provide important and complementary perspectives on models’ ability to explain data. They define a Pareto frontier, where models that rule out more regularities, yet capture the regularities that are present in real data, are preferred. Of course these two dimensions are not the only two that matter for evaluating models, and we do not speak to other important concerns such as parameter estimation and causal inference. Nevertheless, the proposed measures may be relevant to those problems as well: If a model can fit almost any data set, then its good fit to a specific real data set does not necessarily mean that the model is the “right” model for that data. This may suggest caution about how the model parameters should be interpreted.

Our goal is to provide a measure that is easy to apply to a large variety of contexts.

To this end, our measure of restrictiveness is chosen to be computable without the guidance of theoretical results about the model’s implications or empirical content. This differentiates our measure from measures such as the model’s VC dimension, or its hit-rate and accuracy-rate as defined in Selten (1991).¹ (Section 2.3 reviews the related literature and relates it to our work.)

To further clarify the content of our restrictiveness measure, we develop an axiomatization for it in Section 3. The main axioms require that the measure is homogeneous in the unit scale used to quantify model error, and that the measure has a linearity property as the background constraints are varied. An additional “symmetry” axiom requires that the model’s ability to approximate different synthetic data sets has the same effect on the restrictiveness measure. Dropping this axiom returns a broader class of restrictiveness measures, where instead of averaging across synthetic data sets, the data sets are weighted by an analyst’s prior. We develop estimators for both the restrictiveness and completeness measures in Section 4, and prove results about their asymptotic properties, so that users can compute confidence intervals.

The remainder of the paper illustrates the applicability of the restrictiveness measure by using it to evaluate models from three economic domains: (1) predicting certainty equivalents for binary lotteries (where we evaluate *Cumulative Prospect Theory* and *Disappointment Aversion*); (2) predicting initial play in matrix games (where we evaluate the *Poisson Cognitive Hierarchy Model (PCHM)*, *Logit PCHM*, and *Logit Level-1*); and (3) predicting takeup of microfinance in Indian villages (where we evaluate linear regression models based on economically-motivated regressors, and a structural model of diffusion). The first two settings use data from the lab, our third application uses field data. In each of these domains, these measures reveal new insights about the models we examine.

Application 1: Certainty Equivalents. We evaluate two models on a set of binary lotteries from Bruhin et al. (2010): a popular three-parameter specification of Cumulative Prospect Theory (Tversky and Kahneman, 1992), henceforth CPT, and a two-parameter specification of Disappointment Aversion (Gul, 1991), henceforth DA. We find that CPT performs strikingly well on the Bruhin et al. (2010) data, achieving

¹There are representation theorems for many non-parametric theories of individual choice, and some analytic results for the sets of equilibria in games, but we are unaware of representation theorems for the functional forms that are commonly used in applied work.

a completeness of 95%, while DA’s completeness is only 27%.

One explanation for this finding is that CPT is a much better model of risk preferences than DA. Another possibility is that CPT is simply more flexible. We thus evaluate the restrictiveness of the two models, where our background constraints are that the synthetic average certainty equivalents must lie within the range of the lotteries’ possible payoffs, and must respect first-order stochastic dominance. We find that CPT is indeed substantially less restrictive than DA: CPT performs better than DA not only on the real data set but also on the other admissible data sets. Thus, while CPT performs better for predicting the real data, DA rules out more behaviors.

Besides comparing distinct models such as CPT and DA, restrictiveness and completeness can also be compared across nested models to reveal the role played by specific parameters. Adding a parameter always at least weakly increases completeness and decreases restrictiveness, but some parameters achieve greater improvements in completeness for the same decrease in restrictiveness. We find that several parameters lead to large drops in restrictiveness in return for only marginal improvements in completeness, suggesting that these parameters may add flexibility in the wrong directions. The CPT parameter that governs the curvature of the probability weighting function, however, achieves a large improvement in completeness compared to the flexibility it adds, so this parameter seems to capture an important part of risk preferences.

Application 2: Initial Play in Games. Next, we evaluate three models on a set of 3×3 matrix games from Fudenberg and Liang (2019): the Poisson Cognitive Hierarchy Model or *PCHM* (Camerer et al., 2004); *Logit PCHM* (Wright and Leyton-Brown, 2014), which allows for logistic best replies in the PCHM; and *Logit Level-1*, which models the distribution of play as a logistic best reply to the uniform distribution. We impose the background constraint that strictly dominant actions are played at least as often as if by chance (i.e. with probability at least $1/3$) and that strictly dominated actions are played with probability no more than $1/3$.

We find that all three models are highly restrictive, and Logit PCHM and Logit Level-1 are almost twice as complete as PCHM. Moreover, Logit Level-1 and Logit PCHM cannot be statistically distinguished based on either completeness or restrictiveness, suggesting that their empirical content is quite similar. This observation is

not evident from their functional forms, which do not bear an obvious relationship, and involve a different number of free parameters (Logit PCHM uses two, while Logit Level-1 uses only one).

Application 3: Diffusion on a Social Network. Finally, we consider the prediction of microfinance takeup rates in the set of Indian villages studied by Banerjee et al. (2013, 2019), and compare the performance of OLS regression on various economically-motivated regressors with that of an economically-motivated partially linear model built upon “network gossip centrality.” Here we find that the partially linear model is dominated by a simple OLS model based on the average eigenvector centrality of leaders as regressors: the latter has both higher restrictiveness and higher completeness.

Besides these specific findings about each of these economic domains, our analyses make the high-level point that it is not sufficient to count parameters to understand a model’s restrictiveness. Even with just 3 parameters, CPT is not very restrictive on the domain of binary lotteries, and models with different numbers of parameters (such as Logit PCHM and Logit Level-1) turn out to be similarly restrictive. These comparisons are not easy to see from the functional forms associated with a model, but they are revealed by our restrictiveness measure.

2 Approach

2.1 Restrictiveness

Let X be an observable *feature vector* taking values in a finite set \mathcal{X} , and Y be an *outcome variable* taking values in a bounded set $\mathcal{Y} \subseteq \mathbb{R}^k$, $k < \infty$. We use P^* to denote the joint distribution of (X, Y) , P_X^* to denote the marginal distribution of X and $P_{Y|X}^*$ to denote the conditional distribution of Y given X . We assume that the marginal P_X^* is known to the analyst, while the conditional distribution is not.²

The set of all mappings $f : \mathcal{X} \rightarrow \mathcal{Y}$ is denoted $\mathcal{F}^* \equiv \mathcal{Y}^{|\mathcal{X}|}$. We take as a primitive the *discrepancy* function $d : \mathcal{F}^* \times \mathcal{F}^* \rightarrow \mathbb{R}_+$ where $d(f, f')$ tells us how different the two mappings f and f' are. For example, if $\mathcal{Y} \subseteq \mathbb{R}$, then a natural choice for d is

²For example, in a decision theory experiment the experimenter knows the distribution over menus that the subjects will face.

$d(f, f') = \mathbb{E}_{P_X^*}[(f(X) - f'(X))^2]$, i.e., the expected mean-squared distance between the predictions. We allow for functions d that are not distances,³ but we require that $d(f, f') = 0$ if and only if $f = f'$.

Our goal is to evaluate the restrictiveness of a given parametric model $\mathcal{G} = \{g_\theta\}_{\theta \in \Theta} \subseteq \mathcal{F}^*$, where the mappings g_θ are indexed by a finite dimensional parameter θ and Θ is a compact set.⁴ In some cases, it may make sense to evaluate the restrictiveness of the model \mathcal{G} with respect to the set \mathcal{F}^* of all possible mappings from \mathcal{X} to \mathcal{Y} , but in many applications, the analyst has some prior knowledge or intuition about the setting. In these cases, it is often understood that the model obeys some background restrictions, and the question of interest is how restrictive the model is *beyond* these basic background constraints. Formally, we define our measure of restrictiveness for the model \mathcal{G} relative to a set of “admissible” mappings $\mathcal{F} \subseteq \mathcal{F}^*$.

We define the restrictiveness of a model to be the ratio of its expected discrepancy to a mapping f drawn uniformly at random from the admissible set, to the expected discrepancy of a baseline mapping f_{base} chosen to suit the setting.

Definition 1. The restrictiveness of model \mathcal{G} with respect to admissible set \mathcal{F} is

$$r(\mathcal{G}, \mathcal{F}) = \frac{\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}[d(\mathcal{G}, f)]}{\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}[d(f_{\text{base}}, f)]} \quad \forall \mathcal{G}, \mathcal{F} \quad (1)$$

where $\lambda_{\mathcal{F}}$ denotes the uniform distribution on \mathcal{F} , and $d(\mathcal{G}, f) := \min_{g \in \mathcal{G}} d(g, f)$.⁵

This ratio is well-defined as long as the denominator exceed zero, so we will impose this an assumption going forward:

Assumption 1. $\mathbb{E}_{f \sim \lambda_{\mathcal{F}}}[d(f_{\text{base}}, f)] > 0$.

We interpret the performance of f_{base} as a lower bound that any sensible model should outperform.⁶ Normalizing with respect to a baseline has several advantages:

³For example in a subsequent application where \mathcal{Y} is a set of probability distributions, we choose $d(f, f') = \mathbb{E}_{P_X^*}[D(f(X)||f'(X))]$ where D denotes the Kullback-Leibler divergence.

⁴Our subsequent definitions and results do not depend on whether the parametric model \mathcal{G} is point-identified or set-identified.

⁵Since \mathcal{F} is a measurable subset of bounded finite-dimensional Euclidean space, the uniform distribution on \mathcal{F} is well-defined. We discuss a generalization that does not require the uniform distribution in Section 3.

⁶For example, in our application to predicting initial play in games, we define the baseline mapping to predict a uniform distribution of play in every game. Note that while the choice of baseline

First, it makes our measure unitless, so rescaling units does not change the restrictiveness of the model.⁷ Second, whenever the model’s discrepancy to each admissible mapping f is weakly smaller than the discrepancy between the baseline mapping and f (as would be the case whenever f_{base} is chosen from \mathcal{G}), then restrictiveness ranges from an easily interpretable zero to 1. A model with $r = 0$ is completely unrestrictive, while a model with $r = 1$ fits synthetic data no better than the baseline mapping does. If a model performs well on real data and is also highly restrictive, then its good performance occurs not simply because the model can fit any data, but because it precisely identifies regularities in real behavior.

2.2 Completeness

While restrictive models are desirable, a restrictive model is not particularly useful if it rules out the regularities that are present in real data. In our subsequent applications, we thus evaluate models from the dual perspectives of how restrictive they are and how well they fit the actual data. To evaluate model fit, we use the *completeness* measure introduced in Fudenberg et al. (2021): Choose a loss function l so that $l(Y, Y')$ is the error to predicting Y' when the outcome is Y . Then the mapping f^* that minimizes expected loss on the real data is given by

$$f^* \in \arg \min_{f \in \mathcal{F}^*} e_{P^*}(f), \quad \text{where } e_{P^*}(f) := \mathbb{E}_{P^*} [l(f(X), Y)], \forall f \in \mathcal{F}.$$

For example, if \mathcal{X} is a set of lotteries and \mathcal{Y} are subjects’ reported certainty equivalents for each lottery, and l is mean-squared error, then f^* takes each lottery into its average certainty equivalent across subjects. If \mathcal{X} is a set of payoff matrices, \mathcal{Y} is the set of distributions over actions, and $l(Y, Y')$ is Kullback-Leibler divergence from Y' to Y , then f^* maps each game to the corresponding distribution over actions.

mapping affects the value of restrictiveness, it does not affect an ordinal comparison of restrictiveness across two models.

⁷In our first application in Section 5, where we predict certainty equivalents for lotteries, scaling up the lotteries’ payoffs scales up $d(\mathcal{G}, f)$ as well, even though the flexibility of the model has not changed.

Definition 2 (Fudenberg et al., 2021). The *completeness* of model \mathcal{G} is defined by

$$\kappa(\mathcal{G}) := \frac{e_{P^*}(f_{\text{base}}) - \min_{f \in \mathcal{G}} e_{P^*}(f)}{e_{P^*}(f_{\text{base}}) - e_{P^*}(f^*)}.$$

By construction, the measure κ is scale-free and lies within the unit interval. A large completeness κ suggests that the model is able to approximate the real data well: at the extremes, a model with $\kappa = 1$ matches the true mapping f^* exactly, while a model with $\kappa = 0$ is no better at matching f^* than the baseline mapping f_{base} .

Completeness is defined using the loss function l , while restrictiveness is defined using the discrepancy function d . When the discrepancy function d and the loss function l are “paired” in the way described in Online Appendix O.3, then completeness and restrictiveness can be related as follows:

$$\kappa(\mathcal{G}) = 1 - r(\mathcal{G}, \{f^*\}).$$

In other words, completeness is complement of the restrictiveness of model \mathcal{G} with respect to the true model f^* . Our first and third application use mean-squared error as the loss function and expected squared distance as the discrepancy function; our second application uses negative log-likelihood as the loss function and expected KL divergence as the discrepancy function. Both of these are examples of paired functions. We will report both restrictiveness r and completeness κ for each application that we consider.

2.3 Relationship to the Literature

Koopmans and Reiersol (1950) defined a model to be *observationally restrictive* if the distributions it allows are a proper subset of the distributions that would otherwise be possible.⁸ A model that is not observationally restrictive can perfectly match all data and so has $r = 0$.

Selten (1991) proposed a related quantitative measure of flexibility, which has been applied in Beatty and Crawford (2011), Hey (1998), and Harless and Camerer (1994)

⁸As Koopmans and Reiersol (1950) points out, a special case of an observationally restrictive specification is an overidentifying restriction. See e.g. Sargan (1958), Hausman (1978), Hansen (1982), and Chen and Santos (2018) for econometric tests of overidentification.

among others, to understand the restrictiveness of nonparametric economic models (for example, the restrictiveness of the Generalized Axiom of Revealed Preference). The Selten (1991) measure for the flexibility of a model is the fraction of possible data sets that the model can exactly explain. Subtracting this measure from 1 yields a special case of an un-normalized version of our restrictiveness measure, where all mappings are admissible, and the discrepancy function d takes value zero if $f = f'$ and otherwise takes value 1. Our measure allows for constraints on the possible data, and for alternative choices of d that capture approximate fit.⁹ These differences can lead to very different conclusions. For example, consider the set $\{0, 1/n, \dots, (n-1)/n, 1\}$ as a model for the unit interval: This model has measure zero, so it is extremely restrictive according to Selten (1991)'s measure for any value of n . In contrast, according to our measure with the standard mean-squared error discrepancy, the model's restrictiveness decreases in n , and is very unrestrictive for large n .¹⁰

In considering approximate rather than exact fit, our approach is related to papers that measure the distribution of the Afriat index (Choi et al., 2007; Polisson et al., 2020).¹¹ These approaches are motivated by the testing of rationality of choices; our aim here is to show that similar techniques can be applied to a substantially broader class of models. Our use of synthetic data to evaluate restrictiveness is similar in spirit to the use of simulated data to evaluate the power of a hypothesis test, as in Bronars (1987)'s numerical evaluation of a test of GARP proposed by Varian (1982), but our objective is measuring the content of a model's restrictions, and not hypothesis testing.

⁹Beatty and Crawford (2011) propose an alternative “smoothed out” version of Selten (1991)'s measure for the revealed preference setting, which corresponds to another choice of discrepancy d . This measure is closer to our objective of assessing a model's approximate fit.

¹⁰It is also typically difficult to determine whether a parametric model can exactly fit a given data set without the guidance of prior analytical results. For example, Beatty and Crawford (2011) analytically derive the set of budget shares that are consistent with GARP, and Harless and Camerer (1994) use results about generalized expected utility theories to determine whether choices between specially chosen pairs of lotteries (for example, lotteries sharing a common ratio of outcome probabilities) are consistent with those theories. But we do not know how to analytically determine the predictions that are consistent with, say, PCHM or the structural model of microfinance takeup in Application 3.

¹¹Choi et al. (2007) and Polisson et al. (2020) relaxed the implications of expected utility maximization using Afriat's “efficiency index” as an analog of our loss function. They then compare the distribution of the efficiency indices of the actual subjects with the distribution of efficiency indices in randomly generated data.

Our paper is related to a literature in statistics and econometrics on model selection, which dates back to Cox (1961, 1962). There are several key differences. First, classic measures such as AIC and BIC are based on observed data, while our restrictiveness measure is not. This difference reflects a difference in objectives: A primary goal of model selection is to avoid overfitting a complex model to a finite (and small) quantity of data. Thus, these metrics typically trade off some notion of completeness against some notion of restrictiveness: For example, the AIC combines the log-likelihood, which is about fitness to real data (conceptually corresponding to “completeness”) and the number of parameters, which is about the flexibility of the model without reference to real data (conceptually corresponding to “restrictiveness”) in an additive way. In contrast, we view model restrictiveness (and relatedly, model parsimony) as intrinsically valuable even with an infinite data set. Our measure of restrictiveness thus does not depend on sample sizes or fit to real data.¹²

Our measure is also related to complexity measures from computer science and statistics, including the well-known VC dimension, the related Rademacher complexity, and various notions of metric and bracketing entropies. VC dimension is generally nontrivial to derive, and is known for very few economic models.¹³ Rademacher complexity can be numerically computed, but differs from our measure in that it is indexed not only to the model but also to a sample size. (This is again because of the difference in intended uses of the metrics: Rademacher complexity is typically used to bound generalization error, with a view towards avoiding overfitting.) Our restrictiveness measure is instead developed for a conceptually infinite data set: For example, in the analysis of CPT, we do not generate synthetic finite data sets of lotteries and certainty equivalents, but rather directly generate mappings from lotteries to conditional expectations. Thus our definition of restrictiveness makes no reference to the sample size.¹⁴

Various notions of metric entropies (based on the concept of covering numbers) and bracketing entropies (based on the concept of bracketing numbers) have also

¹²Note also that that our notion of “restrictiveness” is distinct from the issue of the identifiability of parameters and the efficiency of estimators.

¹³A recent exception is the work of Basu and Echenique (2020) on the VC dimension of some models of decision-making under uncertainty.

¹⁴We could loosely interpret the proposed restrictiveness measure as analogous to a limiting case of Rademacher complexity for large samples, but we use our discrepancy function d , rather than correlation, to measure the model’s ability to fit the synthetic data.

been used to measure the complexity of a function class in statistics, especially in the theory of empirical processes (Van Der Vaart and Wellner, 1996). Usually, analysts are only concerned with the finiteness and the rates of growth of certain entropy integrals, which are then used to derive bounds on the rates of convergence of various statistical quantities. Relative to these measures, our notion of restrictiveness is different in several respects. First, it is more interpretable, because it provides an explicit quantification of how well the model can approximate a range of possible data. Second, our framework makes it easy to impose economically-motivated background constraints and evaluate the relative restrictiveness of a certain model beyond them. Third, normalization with respect to a baseline makes it easier to compare restrictiveness across different models.

Finally, our paper joins a recent literature on experimental design and the interpretation of experimental results. This includes, for example, DellaVigna and Pope (2019), de Quidt et al. (2018), and Shmaya and Yariv (2016) on the external validity of experimental results; Andrews and Kasy (2019), Frankel and Kasy (2019), and DellaVigna and Linos (2020) on p -hacking and publication bias; Fudenberg and Levine (2020) on the interpretation of natural experiments; and Chassang et al. (2012), Chemla and Hennessy (2019), and Banerjee et al. (2020) on the interpretation of randomized controlled trials.

3 Axiomatic Foundation

Endow the set \mathcal{F}^* of all mappings f with the Lebesgue σ -algebra and a σ -finite measure μ , which can be interpreted as the analyst's prior. We will first define an *approximation error* e , which takes as input the model $\mathcal{G} \subseteq \mathcal{F}^*$, a Lebesgue-measurable set of admissible mappings $\mathcal{F} \subseteq \mathcal{F}^*$, and a discrepancy d . The quantity $e(\mathcal{G}, \mathcal{F}, d)$ is interpreted as the approximation error of the model \mathcal{G} to the admissible set \mathcal{F} , where the quality of the approximation is measured using d . We would like for this approximation error function to satisfy the following axioms.

Axiom 1 (Nonnegativity). For every model \mathcal{G} , admissible set \mathcal{F} , and discrepancy d , $e(\mathcal{G}, \mathcal{F}, d) \geq 0$.

Axiom 2 (Monotonicity). Fix any set of admissible mappings \mathcal{F} . If the sets \mathcal{G}_1 and

\mathcal{G}_2 satisfy $d(\mathcal{G}_1, f) > d(\mathcal{G}_2, f)$ for all $f \in \mathcal{F}$, then $e(\mathcal{G}_1, \mathcal{F}, d) > e(\mathcal{G}_2, \mathcal{F}, d)$. And if $d(\mathcal{G}_1, f) = d(\mathcal{G}_2, f)$ for all $f \in \mathcal{F}$ then $e(\mathcal{G}_1, \mathcal{F}, d) = e(\mathcal{G}_2, \mathcal{F}, d)$.

Axiom 2 says that if one model is better able to approximate every admissible mapping than another, the first model has lower approximation error.

Axiom 3 (Rescaling of Units). (a) Fix any model \mathcal{G} , set of admissible mappings \mathcal{F} , and discrepancy d . Then $e(\mathcal{G}, \mathcal{F}, \alpha \cdot d) = \alpha \cdot e(\mathcal{G}, \mathcal{F}, d)$ for every $\alpha \in \mathbb{R}_+$

(b) Fix any set of admissible mappings \mathcal{F} and discrepancy d . If \mathcal{G}_1 and \mathcal{G}_2 satisfy $d(\mathcal{G}_1, f) = \alpha \cdot d(\mathcal{G}_2, f)$ for all $f \in \mathcal{F}$, then $e(\mathcal{G}_1, \mathcal{F}, d) = e(\mathcal{G}_2, \mathcal{F}, \alpha \cdot d)$.

Part (a) of Axiom 3 says that any rescaling of the units of the discrepancy d is inherited also by the approximation error measure. Part (b) says that scaling the discrepancy between a model \mathcal{G} to each mapping f leads to the same value of approximation error as scaling the units of the discrepancy d .

Axiom 4 (Linearity). For any countable sequence of disjoint measurable sets $\mathcal{F}_1, \mathcal{F}_2, \dots$ whose union $\mathcal{F} \equiv \cup_{i=1}^{\infty} \mathcal{F}_i$ has strictly positive measure,

$$e(\mathcal{G}, \mathcal{F}, d) = \sum_{i=1}^{\infty} \frac{\mu(\mathcal{F}_i)}{\mu(\mathcal{F})} \cdot e(\mathcal{G}, \mathcal{F}_i, d) \quad \forall \mathcal{G}, d.$$

Consider constraining the set of admissible mappings \mathcal{F} to a subset \mathcal{F}_1 or its complement \mathcal{F}_2 . The *ex post* approximation errors of a model \mathcal{G} with respect to either of these new admissible sets is, respectively, $e(\mathcal{G}, \mathcal{F}_1, d)$ or $e(\mathcal{G}, \mathcal{F}_2, d)$. Axiom 4 says that the *ex ante* approximation error $e(\mathcal{G}, \mathcal{F}, d)$ is a convex combination of the *ex post* approximation errors, where each *ex post* subset contributes to the *ex ante* approximation error in proportion to its measure.

Axiom 5 (Symmetry). Fix any admissible set \mathcal{F} and any bijection τ from \mathcal{F} to itself. Consider two sets \mathcal{G}_1 and \mathcal{G}_2 where

$$d(\mathcal{G}_1, f) = d(\mathcal{G}_2, \tau(f)) \quad \forall f \in \mathcal{F}.$$

Then $e(\mathcal{G}_1, \mathcal{F}, d) = e(\mathcal{G}_2, \mathcal{F}, d)$.

Axiom 5 says that permuting the various discrepancies between the model and the admissible mappings f does not affect the overall approximation error. This reflects a “principle of indifference” over the admissible mappings.

Proposition 1. *An approximation error e satisfies Axioms 1-4 if and only if there is a function $c : \mathcal{F}^* \rightarrow \mathbb{R}$ such that*

$$e(\mathcal{G}, \mathcal{F}, d) = \mathbb{E}_{f \sim \mu_{\mathcal{F}}} \left[c(f) \cdot \min_{g \in \mathcal{G}} d(g, f) \right] \quad \forall \mathcal{G}, \mathcal{F}, d \quad (2)$$

where $\mu_{\mathcal{F}}$ denotes the measure μ conditional on the event \mathcal{F} . If additionally e satisfies Axiom 5, then

$$e(\mathcal{G}, \mathcal{F}, d) = \mathbb{E}_{f \sim \lambda_{\mathcal{F}}} \left[\min_{g \in \mathcal{G}} c \cdot d(g, f) \right] \quad \forall \mathcal{G}, \mathcal{F}, d \quad (3)$$

for a positive constant c , where λ denotes the Lebesgue measure on \mathcal{F}^* .

Our restrictiveness measure assumes (3), and normalizes the approximation error of model \mathcal{G} relative to the approximation error of the baseline mapping f_{base} .

The proof of Proposition 1 (and all other results in this paper) can be found in the appendix, but we briefly sketch our proof strategy. It is clear that Axioms 1-4 are satisfied by the representation in (2), and that these axioms and Axiom 5 are satisfied by the representation in (3), so necessity is immediate. To demonstrate sufficiency, we first fix a model \mathcal{G} and a discrepancy d , and define $e_{\mathcal{G},d}(\mathcal{F}) \equiv e(\mathcal{G}, \mathcal{F}, d)$. A1 and A4 together imply that the function $\nu : \Sigma \rightarrow \mathbb{R}$, which satisfies $\nu(\mathcal{F}) = e_{\mathcal{G},d}(\mathcal{F}) \cdot \mu(\mathcal{F})$ for every $\mathcal{F} \in \Sigma$, is a measure. We use this observation to show that

$$e_{\mathcal{G},d}(\mathcal{F}) = \mathbb{E} [h_{\mathcal{G},d}(f) : f \sim \mu_{\mathcal{F}}] \quad \forall \text{measurable } \mathcal{F}$$

where $h_{\mathcal{G},d}$ is the Radon-Nikodym derivative of ν with respect to μ . We then use A2 and A3 to show that $h_{\mathcal{G},d}(f)$ must be linear in the distance between the model \mathcal{G} and the mapping f , yielding (2). Finally, further requiring equivalence up to permutations of the mappings (A5) yields the representation in (3).

Dropping Axiom 5 returns a broader class of restrictiveness measures, where the weight placed on the model's discrepancy can differ across admissible mappings. We impose Axiom 5 throughout the rest of this paper, thus assuming a uniform distribution over the admissible set. Besides its computational advantages,¹⁵ and its

¹⁵The uniform distribution is easy to implement even for admissible sets \mathcal{F} with potentially complicated structures. For example, in our application to prediction of certainty equivalents, we

motivation via Axiom 5, there are a number of conceptual justifications for this distribution. First, once the admissible set \mathcal{F} is specified, the uniform distribution on \mathcal{F} is pinned down. This reduces the number of primitives to be chosen, and helps prevent cherry-picking with respect to the distribution on \mathcal{F} . Second, under the uniform distribution, (measurable) subsets of mappings in \mathcal{F} are sampled proportionally to the Lebesgue measure, which is the most standard measure of size. In this regard, our use of the uniform distribution is parallel to Selten (1991)'s use of area (see Section 2.3). Finally, under the uniform distribution, the interpretation of r is simply the normalized average discrepancy over the space \mathcal{F} of admissible mappings. Non-uniform weightings over \mathcal{F} require additional specification, and make the dependence of restrictiveness on \mathcal{F} less transparent.¹⁶

4 Estimates and Test Statistics

We now discuss how to implement our approach in practice. Recall that we restrict \mathcal{X} to be finite, so \mathcal{F}^* is finite-dimensional. In Appendix O.4, we provide a discussion on how to compute restrictiveness and estimate completeness when \mathcal{X} is a continuum and \mathcal{F}^* is infinite-dimensional.

4.1 Computing Restrictiveness r

We provide an algorithm for computing r : Randomly sample M times from the admissible set \mathcal{F} , and for each sampled $f_m \in \mathcal{F}$, compute $d(\mathcal{G}, f_m)$ and $d(f_{\text{base}}, f_m)$. Then

$$\hat{r} := \frac{\frac{1}{M} \sum_{m=1}^M d(\mathcal{G}, f_m)}{\frac{1}{M} \sum_{m=1}^M d(f_{\text{base}}, f_m)}$$

is an estimator for restrictiveness $r = r(\mathcal{G}, \mathcal{F})$. In principle, the number of simulations we run, M , can be taken as large as we want, so \hat{r}_M can be made arbitrarily close to r by the Law of Large Numbers.

build monotonicity with respect to FOSD into our definition of \mathcal{F} , and it is straightforward to sample uniformly from \mathcal{F} by first sampling from a larger space without the monotonicity constraints, and then only keeping the draws that satisfy the monotonicity constraints.

¹⁶Note that assigning weights over \mathcal{F} based on some notion of distance or difference from the real mapping f^* would make the comparison of restrictiveness of two models dependent on the real data, blurring its distinction with restrictiveness.

Moreover, the approximation error under a given finite M can be quantified based on the standard Central Limit Theorem and the Delta Method:

Proposition 2. *Under Assumption 1,*

$$\frac{\sqrt{M}(\hat{r}_M - r)}{\hat{\sigma}_{\hat{r}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

where asymptotic variance estimator $\hat{\sigma}_{\hat{r}}^2$ is defined by

$$\hat{\sigma}_{\hat{r}}^2 := \frac{\hat{\sigma}_{\mathcal{G}}^2 - 2 \cdot \hat{r} \cdot \hat{\sigma}_{\mathcal{G}, f_{base}} + \hat{r}^2 \cdot \hat{\sigma}_{f_{base}}^2}{\left(\frac{1}{M} \sum_{m=1}^M d(f_{base}, f_m)\right)^2}, \quad (4)$$

with $\hat{\sigma}_{\mathcal{G}}^2$ being the sample variance of $d(\mathcal{G}, f_m)$, $\hat{\sigma}_{f_{base}}^2$ the sample variance of $d(f_{base}, f_m)$, and $\hat{\sigma}_{\mathcal{G}, f_{base}}^2$ the sample covariance of $d(\mathcal{G}, f_m)$ and $d(f_{base}, f_m)$, across $m = 1, \dots, M$.

Confidence intervals for r can also be constructed in standard ways. We again note that the confidence intervals here simply measure the approximation error of r based on a finite number of simulations and do not reflect randomness in experimental data.

4.2 Estimating Completeness κ

Suppose that the analyst has access to a finite sample of data $\{Z_i := (X_i, Y_i)\}_{i=1}^N$ drawn from the unknown true distribution P^* . To estimate completeness, which is defined based on the loss function l introduced in Section 2.2, we use K -fold cross-validation to estimate the out-of-sample prediction error of the model.¹⁷ (In our applications, we take the standard choice of $K = 10$.) Specifically, we randomly divide \mathbf{Z}_N into K (approximately) equal-sized groups. To simplify notation, assume that $J_N = \frac{N}{K}$ is an integer. Let $k(i)$ denote the group number of observation Z_i , and fix an arbitrary set of mappings $\tilde{\mathcal{F}}$. For each group $k = 1, \dots, K$, define

$$\hat{f}^{-k} := \arg \min_{f \in \tilde{\mathcal{F}}} \frac{1}{N - J_N} \sum_{k(i) \neq k} l(f, Z_i)$$

¹⁷Alternatively, we can use the in-sample error estimator without cross validation, if we are not concerned with out-of-sample errors in finite samples. See, e.g., the estimator in Appendix O.4.

to be the element of $\tilde{\mathcal{F}}$ that minimizes error for prediction of observations outside of group k . This estimated mapping is used for prediction of the k -th test set, and

$$\hat{e}_k := \frac{1}{J_N} \sum_{k(i)=k} l(\hat{f}^{-k}, Z_i)$$

is its out-of-sample error on the k -th test set. Then,

$$\hat{e}_{CV}(\tilde{\mathcal{F}}) := \frac{1}{K} \sum_{k=1}^K \hat{e}_k$$

is the average test error across the K folds. This is an estimator for the unobservable expected error of the best mapping from class $\tilde{\mathcal{F}}$.

Setting $\tilde{\mathcal{F}}$ to be \mathcal{F}^* , \mathcal{G} , or $\{f_{\text{base}}\}$, we can compute $\hat{e}_{CV}(\mathcal{F}^*)$, $\hat{e}_{CV}(\mathcal{G})$ and $\hat{e}_{CV}(f_{\text{base}})$ from the data, leading to the following estimator for κ :

$$\hat{\kappa} = 1 - \frac{\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}^*)}{\hat{e}_{CV}(f_{\text{base}}) - \hat{e}_{CV}(\mathcal{F}^*)}.$$

It is crucial that the denominator in $\hat{\kappa}$ does not vanish asymptotically, so we impose the following assumption:

Assumption 2 (Baseline Mapping is Imperfect). $e_{P^*}(f_{\text{base}}) - e_{P^*}(f^*) > 0$.

This assumption says that the baseline mapping performs strictly worse in expectation than the best mapping so there is some room for a model to do better. Under additional technical conditions we show, by applying and adapting Proposition 5 in Austern and Zhou (2020), that $\hat{\kappa}$ is asymptotically normal.

Proposition 3. *Under Assumption 2 and some regularity conditions,*¹⁸

$$\frac{\sqrt{N}(\hat{\kappa} - \kappa)}{\hat{\sigma}_{\hat{\kappa}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where the estimate $\hat{\sigma}_{\hat{\kappa}}$ of the standard error is as defined in Appendix B.2.

¹⁸See Appendix B for details of these assumptions.

5 Application 1: Certainty Equivalents

5.1 Setting

Our first application is to the prediction of certainty equivalents for a set of 25 binary lotteries from Bruhin et al. (2010). Each lottery is described as a tuple $x = (\bar{z}, \underline{z}, p)$, where $\bar{z} > \underline{z} \geq 0$ are the possible prizes, and p is the probability of the larger prize. Each observation is a pair consisting of a lottery and a reported certainty equivalent by a given subject, so we can describe the feature space \mathcal{X} by the 25 lottery tuples $(\bar{z}, \underline{z}, p)$ in the Bruhin et al. (2010) data, and the outcome space by $\mathcal{Y} = \mathbb{R}$. Note that the residual uncertainty in Y conditional on X reflects heterogeneity in certainty equivalents reported across subjects for the same lottery.¹⁹

We predict the average certainty equivalent (over subjects) for each lottery in this data set. A mapping for this problem is any function $f : \mathcal{X} \rightarrow \mathbb{R}$ from the 25 lotteries to average certainty equivalents, and the discrepancy $d(f, f')$ between two mappings f and f' is defined to be the average mean-squared distance between the two mappings' predictions

$$d(f, f') = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} (f(x) - f'(x))^2.$$

We evaluate the restrictiveness and completeness of two economic models. First we consider a three-parameter version of *Cumulative Prospect Theory* indexed by $\theta = (\alpha, \gamma, \delta)$, which specifies a “utility”

$$w(p)v(\bar{z}) + (1 - w(p))v(\underline{z}) \tag{5}$$

for each lottery $(\bar{z}, \underline{z}, p)$, where

$$v(z) = z^\alpha \tag{6}$$

is a value function for money, and

$$w(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma} \tag{7}$$

¹⁹In Appendix O.2, we discuss how to extend our approach to allow for subject-level heterogeneity.

is a probability weighting function.²⁰ The predicted certainty equivalent of a binary lottery is

$$g_\theta(\bar{z}, \underline{z}, p) = v^{-1}(w(p)v(\bar{z}) + (1 - w(p))v(\underline{z})).$$

Following the literature, we restrict $\alpha, \gamma \in [0, 1]$, and $\delta \geq 0$. We specify \mathcal{G} as the set of all such functions g_θ with parameters θ in this range, and refer to this model simply as CPT. As a baseline, we consider the function f_{base} that maps each lottery into its expected value, corresponding to $\alpha = \gamma = \delta = 1$.

Second, we consider the *Disappointment Aversion* model of Gul (1991), using a parametric form proposed in Routledge and Zin (2010) with the parameters $\lambda = (\alpha, \eta)$, where $\alpha \in [0, 1]$ and $\eta > -1$.²¹ The value function for money is the same as in (6), but the probability weighting function is given instead by

$$\tilde{w}(p) = \frac{p}{1 + (1 - p)\eta}$$

There are two parameters: α again reflects the curvature of the utility function, while $\eta > 0$ corresponds to “disappointment aversion,” namely aversion to realizations of the lottery that are worse than its certainty equivalent.

The predicted certainty equivalent is

$$g_\lambda(\bar{z}, \underline{z}, p) = v^{-1}(\tilde{w}(p)v(\bar{z}) + (1 - \tilde{w}(p))v(\underline{z})).$$

We specify \mathcal{G}_λ as the set of all such functions g_λ , and refer to this model simply as DA. Again, we use as a baseline the expected value mapping, which corresponds to $\alpha = 1$ and $\eta = 0$ in DA.

5.2 Completeness

CPT achieves a striking out-of-sample performance for predicting the average certain equivalents in Bruhin et al. (2010) data: it is 95% complete. (A similar result was reported in Fudenberg et al. (2021) for a pooled sample of gain-domain and loss-

²⁰This parametric form for $w(p)$ was first suggested by Goldstein and Einhorn (1987) and Lattimore et al. (1992).

²¹To facilitate comparison with CPT, we depart slightly from Routledge and Zin (2010) by imposing the functional form $v(z) = z^\alpha$ instead of $v(z) = z^\alpha/\alpha$.

domain lotteries.) Thus, the model achieves almost all of the possible improvement in prediction accuracy over the baseline.²² In contrast, DA is only 27% complete on the same data. One explanation is that CPT more precisely captures the observed risk preferences in the data than DA, but another possibility is that CPT is flexible enough to mimic most functions from binary lotteries to certainty equivalents, while DA imposes more substantial restrictions. These explanations have very different implications for how to interpret CPT’s empirical success compared to DA’s.

5.3 Restrictiveness

To distinguish between these explanations, we now compute the restrictiveness of the two models. We define the admissible set to be all mappings satisfying the following criteria:

1. $\underline{z} \leq f(\bar{z}, \underline{z}, p) \leq \bar{z}$
2. if $\bar{z} > \bar{z}'$, $\underline{z} > \underline{z}'$, and $p \geq p'$ then $f(\bar{z}, \underline{z}, p) > f(\bar{z}', \underline{z}', p')$
3. if $\bar{z} \geq \underline{z}$, $p > p'$, then $f(\bar{z}, \underline{z}, p) > f(\bar{z}, \underline{z}, p')$

Constraint (1) requires that the certainty equivalent is within the range of the possible payoffs, while constraints (2) and (3) are equivalent to first-order stochastic dominance.²³

Table 1 reports the completeness and restrictiveness of both models.

	# Param	Restrictiveness	Completeness
CPT	3	0.28 (0.003)	0.95 (0.02)
DA	2	0.47 (0.006)	0.27 (0.06)

Table 1: Completeness for both models is estimated on the real data, which includes reported certainty equivalents by each of 179 subjects. Standard errors for the completeness estimates are computed using a block bootstrapping procedure that clusters together all observations from the same subjects, see Appendix C. Restrictiveness is estimated from 1000 simulations, and we report analytic standard errors using Proposition 2.

²²This finding is consistent with Peysakhovich and Naecker (2017)’s result that CPT approximates the predictive performance of lasso regression trained on a high-dimensional set of features.

²³There are many pairs of lotteries in the Bruhin et al. (2010) lottery data that can be compared via (2) and (3), so these conditions are not vacuous.

The restrictiveness of CPT is 0.28, so on average, CPT’s approximation error is about one fourth of the error of the expected value mapping. DA is more restrictive, with an average approximation error almost one half of the error of the baseline mapping. Thus the two models are not directly comparable: CPT performs substantially better for predicting the real data, but would have performed well out-of-sample given sufficient data from almost any underlying data-generating process that respects first-order stochastic dominance. DA rules out more behaviors that satisfy first-order stochastic dominance, but in doing so is unable to well approximate the actual Bruhin et al. (2010) data.

5.4 The Role of a Parameter

In addition to comparing distinct models such as CPT and DA, our approach can also be used to compare nested models in order to reveal the role played by specific parameters. Adding a free parameter must at least weakly decrease restrictiveness and increase completeness, but we find that parameters can differ substantially in their effectiveness in trading off between these two goals. We also show that models with the same number of parameters can have very different levels of restrictiveness, and thus a simple parameter count is substantively less informative than our measure.

Specifically, we now consider alternative specifications of CPT and DA with fewer free parameters. Some of these specifications have been studied in the literature: CPT(α, γ), with δ set to 1, is the specification used in Karmarkar (1978)²⁴; CPT(γ, δ), with $\alpha = 1$, corresponds to a risk-neutral CPT agent whose utility function over money is $u(z) = z$ but exhibits nonlinear probability weighting; CPT(α), with $\delta = \gamma = 1$, corresponds to an Expected Utility decision-maker whose utility function is as given in (6), and is also equivalent to DA(α).²⁵ The model CPT(γ), with $\alpha = \delta = 1$, and CPT(δ), with $\alpha = \gamma = 1$ have not been studied in the prior literature, but we report them for comparison. We also consider DA(η) as in Gul (1991), with $\alpha = 1$, which corresponds to a disappointment-averse decision maker whose utility is linear in money.

²⁴This specification with weighting function $w(p) = \frac{p^\gamma}{p^\gamma + (1-p)^\gamma}$ is very similar to one used in Tversky and Kahneman (1992), where the weighting function was $w(p) = \frac{p^\gamma}{p^\gamma + (1-p)^\gamma}^{1/\gamma}$.

²⁵See the survey Fehr-Duda and Epper (2012) for further discussion of these different parametric

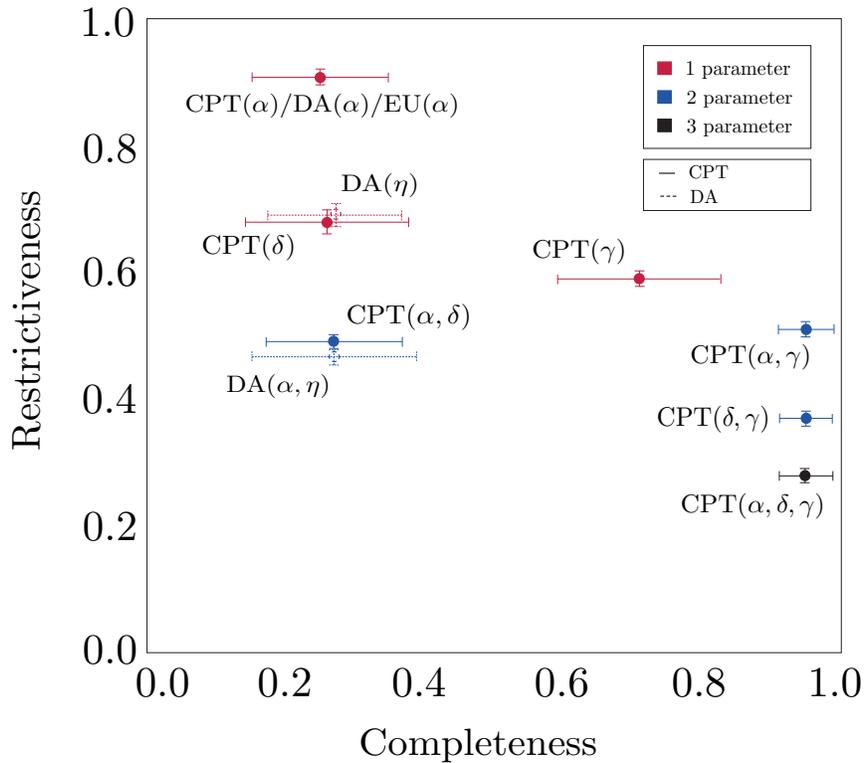


Figure 1: Comparison of models by their completeness and restrictiveness.

Figure 1 plots restrictiveness and completeness for these alternative specifications (see also Table 6 in the appendix). This figure reveals that some specifications fall in the interior of a restrictiveness-completeness Pareto frontier: Each of $CPT(\delta)$, $CPT(\alpha, \delta)$, and $DA(\alpha, \eta)$ are dominated, in the sense that another model is simultaneously more complete and also more restrictive.²⁶ The figure also reveals substantial dispersion in the restrictiveness of these specifications (ranging from $r = 0.28$ to $r = 0.92$), even though all of the specifications use only a small number of parameters. This observation emphasizes the distinction between our method and a simple parameter count.

By looking more specifically at how restrictiveness and completeness vary across

forms, and others which have been proposed in the literature.

²⁶ $CPT(\delta)$ is simultaneously less complete and less restrictive than $DA(\eta)$, while both $CPT(\alpha, \delta)$ and $DA(\alpha, \eta)$ are less complete and less restrictive than the single parameter model $CPT(\gamma)$.

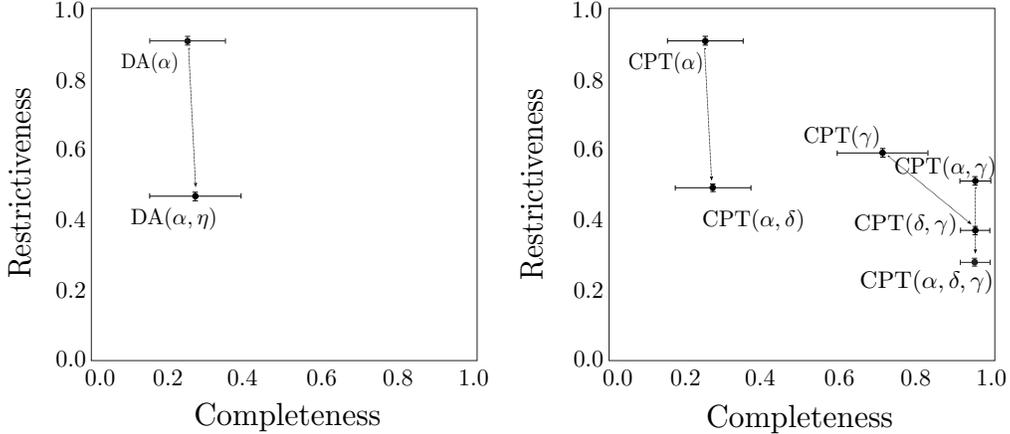
two nested specifications, we can better understand the role that any specific parameter plays. Figure 2 shows that the different parameters used for probability weighting are not equally effective. Adding the parameter δ , which governs the elevation of the probability weighting curve, to any specification of CPT leads to a large drop in restrictiveness in return for only a small gain in completeness. We find a similar result for the “disappointment aversion” parameter η in DA, which barely improves upon the completeness of $DA(\alpha)$, but leads to substantial decrease in restrictiveness. In contrast, the parameter γ , which governs the curvature of the probability weighting function in CPT, appears to play an important role in capturing real risk preferences: Adding γ to any CPT specification leads to a sizeable improvement in completeness at the cost of a modest reduction in restrictiveness. This supports previous findings that that probability distortions play an important role in fitting experimental and field data (Snowberg and Wolfers, 2010; Fehr-Duda and Epper, 2012; Barseghyan et al., 2013), and adds a new perspective by comparing gains in completeness with loss of restrictiveness.

5.5 Robustness Checks

We next investigate how restrictiveness changes as we vary various primitives, and show that our qualitative findings are robust to these variations.

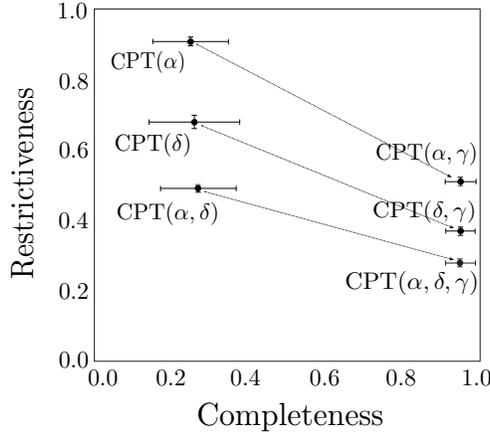
Different distribution over the admissible set. The uniform distribution is the same as $\text{beta}(1, 1)$, so to test the sensitivity of the restrictiveness measure we consider nearby $\text{beta}(a, b)$ distributions, with parameters (a, b) sampled from a uniform distribution over $[0.9, 1.1] \times [0.9, 1.1]$. For each (a, b) pair, we generate certainty equivalents from a $\text{beta}(a, b)$ distribution over the prize range, again keeping only those functions f that satisfy FOSD. Over 100 such distributions $\text{beta}(a, b)$, the average restrictiveness is 0.29, with a minimum value of 0.27 and a maximum value of 0.32.

Different admissible set. Next, we compute the restrictiveness of the model $CPT(\alpha, \delta, \gamma)$ with respect to an admissible set that imposes the range restriction in (1) but drops the FOSD restrictions in (2) and (3). The model’s errors are substantially higher when we drop FOSD (increasing from 63.75 to 102.41), but so are



(a) Role of η in DA

(b) Role of δ in CPT



(c) Role of γ in CPT

Figure 2: Impact of the probability weighting parameters on completeness and restrictiveness.

the errors of the Expected Value benchmark. We find that $CPT(\alpha, \delta, \gamma)$'s relative performance compared to the expected-value baseline is nearly identical regardless of whether we impose FOSD or not: the model's restrictiveness relative to this larger admissible set is 0.29 (compared to a restrictiveness of 0.28 relative to the original admissible set).

Lotteries over the loss domain. We now investigate the performance of CPT with respect to 25 binary lotteries over the loss domain from Bruhin et al. (2010). On this data, the 3-parameter specification of CPT indexed to (β, γ, δ) predicts the

certainty equivalent

$$v^{-1}((1 - w(1 - p)) \cdot v(\bar{z}) + w(1 - p) \cdot v(\underline{z}))$$

for each lottery $(\bar{z}, \underline{z}, p)$, where $v(z) = -((-z)^\beta)$ and $w(p) = (\delta p^\gamma)/(\delta p^\gamma + (1 - p)^\gamma)$. Table 2 reports completeness and restrictiveness measures for the four specifications of CPT that have appeared in the prior literature. The results are very similar to what we found for the gain domain.

	# Param	Restrictiveness	Completeness
CPT(β, γ, δ)	3	0.31 (0.02)	0.92 (0.03)
CPT(β, γ)	2	0.39 (0.02)	0.67 (0.08)
CPT(δ, γ)	2	0.32 (0.01)	0.92 (0.02)
CPT(β)	1	0.83 (0.03)	0.01 (0.09)

Table 2: Restrictiveness and completeness for CPT on lotteries over the loss domain.

Three-outcome lotteries. Finally we evaluate the restrictiveness of CPT(α, δ, γ) on a set of 18 three-outcome gain-domain lotteries from Bernheim and Sprenger (2020). For each lottery $(z_1, z_2, z_3; p_1, p_2, p_3)$, where $z_1 \geq z_2 \geq z_3 \geq 0$, the predicted certainty equivalent is

$$v^{-1}(v(z_1) + w(p_2 + p_3)(v(z_2) - v(z_1)) + w(p_3)(v(z_3) - v(z_2)))$$

where v and w have the same functional forms as used above.

The restrictiveness of CPT on this set of three-outcome lotteries is 0.57, with a standard error of 0.02. Thus CPT is about twice as restrictive on three-outcome lotteries than on binary lotteries.

6 Application 2: The Distribution of Initial Play

6.1 Setting

Our second application is to predicting the distribution of initial play in games. Here the feature space \mathcal{X} consists of the 466 unique 3×3 payoff matrices from Fudenberg and Liang (2019). The outcome space is the set $\mathcal{Y} = \Delta(\{a_1, a_2, a_3\})$ of distributions of row player actions chosen by the participants in the experiments. The analyst seeks to predict this distribution for each game.

For any two mappings f and f' , we define $d(f, f')$ to be the average Kullback-Liebler divergence between the predicted distributions:

$$d(f, f') = \frac{1}{466} \sum_{x \in \mathcal{X}} D(f(x) \| f'(x))$$

where D denotes the Kullback-Leibler divergence.

We consider three economic models: The *Poisson Cognitive Hierarchy Model* (PCHM) of Camerer et al. (2004), the Level-1 model with logistic best replies (henceforth *Logit Level-1*), and the PCHM with logistic best replies (henceforth *Logit PCHM*). The PCHM supposes that there is a distribution over players of differing levels of sophistication: The *level-0* player randomizes uniformly over his available actions, the *level-1* player best responds to level-0 play (Stahl and Wilson, 1994, 1995; Nagel, 1995); and for $k \geq 2$, level- k players best respond to a perceived distribution

$$p_k(h, \tau) = \frac{\pi_\tau(h)}{\sum_{l=0}^{k-1} \pi_\tau(l)} \quad \forall h \in \mathbb{N}_{<k} \quad (8)$$

over (lower) opponent levels, where π_τ is the Poisson distribution with rate parameter $\tau \geq 0$. The parameter τ is the single free parameter of the model.

The *Logit Level-1* prediction is defined as follows. For each row player action a_i , let $\bar{u}(a_i)$ be the expected payoff of a_i when the column player uses a uniform distribution. The predicted frequency with which a_i is played is

$$\frac{\exp(\lambda \cdot \bar{u}(a_i))}{\sum_{i=1}^3 \exp(\lambda \cdot \bar{u}(a_i))}$$

where the logit parameter $\lambda \in \mathbb{R}_+$ is the single free parameter of the model.

The *Logit PCHM* (see e.g. Wright and Leyton-Brown (2014)) replaces the assumption of exact maximization in the PCHM with a logit best response. That is, the level-0 player chooses $g_0 = (1/3, 1/3, 1/3)$ as in the PCHM, but we recursively construct the distribution of play for higher levels as follows. For each $k \geq 1$, define

$$v_k(a_i) = \sum_{h=0}^{k-1} p_k(h, \tau) \left(\sum_{j=1}^3 g_h(a_j) u(a_i, a_j) \right)$$

to be the expected payoff of action a_i against a player whose type is distributed according to $p_k(\cdot, \tau)$, where $p_k(h, \tau)$ is as given in (8). The distribution of play for a level- k player is then

$$g_k(a_i) = \frac{\exp(\lambda \cdot v_k(a_i))}{\sum_{j=1}^3 \exp(\lambda \cdot v_k(a_j))}$$

where $\lambda \in \mathbb{R}_+$ is a logit parameter. We aggregate across levels using a Poisson distribution with rate parameter $\tau \in \mathbb{R}_+$ to yield the predicted distribution of play.

Finally, we define the baseline mapping f_{base} to predict uniform play in every game x . This mapping is nested in all three models.²⁷

6.2 Completeness

The models PCHM, Logit Level-1, and Logit PCHM are 43.6%, 72.7%, and 72.9% complete. Thus, as observed in a related study by Wright and Leyton-Brown (2014), Logit PCHM provides much better predictions of the distribution of play than the baseline PCHM does.

Perhaps surprisingly, almost all of Logit PCHM's improved performance can be obtained by simply adding the logit parameter to the Level-1 model; the further improvement from allowing for multiple levels of sophistication is negligible. Fudenberg and Liang (2019) found that the Level-1 model provides a good prediction of the modal action, but it is not obvious from the previous result that Logit Level-1 would perform so well in predicting the full distribution of play. The fact that it does further suggests that initial play in many of these experiments is rather unstrategic.²⁸

²⁷Let $\tau = 0$ in the PCHM or Logit PCHM, and let $\lambda = 0$ in Logit Level-1.

²⁸Fudenberg and Liang (2019) found that modal play in some sorts of games is better described by

6.3 Restrictiveness

We turn now to evaluating the restrictiveness of these models. Compared to the case of preferences over binary lotteries, economic theory provides very little in the way of a priori restrictions on initial play.²⁹ We thus define the admissible set to include all mappings satisfying the following weak conditions:³⁰

1. If an action is strictly dominated, then the frequency with which it is chosen does not exceed $1/3$.
2. If an action is strictly dominant, then the frequency with which it is chosen is at least $1/3$.

All three models are very restrictive relative to this admissible set: Logit Level-1's restrictiveness is 0.970, PCHM's restrictiveness is 0.992, and Logit PCHM's restrictiveness is 0.971. Since the completeness of these models ranges from 0.436 to 0.729, these models are much better predictors of the real data than of the synthetic data sets.

Table 6.3 reports completeness and restrictiveness measures for all three models. We find that Logit Level-1 and Logit PCHM are substantially more complete than PCHM and only slightly less restrictive. Moreover, Logit Level-1 and Logit PCHM are almost identical in terms of completeness and restrictiveness, even though the parametric forms of the two models are not evidently related.³¹ In the subsequent Online Appendix O.5, we investigate the relationship between Logit Level-1 and Logit PCHM further by studying the correlation in their errors.

Finally, as a robustness check, we consider a strengthening of the background constraints imposed on the admissible set \mathcal{F} . For each $t \in [0, 0.3)$, we define the

equilibrium notions than level-1. Since such regularities cannot be accommodated by the logit level-1 model, these may explain the gap between the completeness of logit level-1 and full completeness. Costa-Gomes et al. (2001) find a sizable fraction of level-2 players in their experimental data, which may further help to explain this gap.

²⁹Payoff maximization implies that dominant strategies should have probability 1 and dominated strategies have probability 0, but this is inconsistent with observed play in most game theory experiments).

³⁰In our data, the median frequency for a strictly dominated action is 0.03, and the highest frequency is 0.35; the median frequency for a strictly dominant action is 0.86, and the lowest frequency is 0.69.

³¹No value of τ in the PCHM yields the Level-1 model, so Logit Level-1 is not nested within Logit PCHM.

Table 3: Restrictiveness and Completeness for Initial Play

	# Param	Restrictiveness	Completeness
PCHM	1	0.992 (<0.001)	0.436 (0.017)
logit level-1	1	0.970 (<0.001)	0.727 (0.015)
logit PCHM	2	0.971 (0.003)	0.729 (0.014)

Restrictiveness is estimated from 1000 simulations.

admissible set $\mathcal{F}(t)$ to include all mappings f that satisfy the following conditions: (1) If an action is strictly dominated, then the frequency with which it is chosen does not exceed $1/3 - t$; (2) If an action is strictly dominant, then the frequency with which it is chosen is at least $1/3 + t$.

The constraint imposed by these conditions increases in t , and $t = 0$ returns our original specification of \mathcal{F} . We find that across choices of $t \in [0, 0.3)$, the restrictivenesses of PCHM, Logit PCHM, and Logit Level-1 do not fall below 0.89 (see Table 4 below). This suggests that the high restrictiveness of these models is robust to constraints imposed only on strictly dominant and strictly dominated actions.

	PCHM	Logit Level-1	Logit PCHM
max	0.993	0.969	0.972
min	0.974	0.890	0.957

Table 4: Largest and smallest restrictiveness measures as t varies over $[0, 0.3)$.

7 Application 3: Diffusion in Social Networks

7.1 Setting

Our final application is to the prediction of microfinance takeup rates following diffusion of information in social networks. We use data from an intervention studied by Banerjee, Chandrasekhar, Duflo, and Jackson (2013), in which certain “leaders” in 43 villages in Karnataka, India were given information about a microfinance program,

and takeup of the program was then tracked.³²

For each village i , let y_i be the average takeup rate among non-leader households.³³ Our goal is to predict y_i given the observed characteristics X_i of village i . Specifically, a village configuration $X_i := (N_i, A_i, L_i)$ consists of a set N_i of villagers, an $n_i \times n_i$ adjacency matrix A_i that represents the measured social network, and finally a set L_i of leaders in village i . The feature space is the collection of 43 village configurations, and we are interested in mappings of the form $f : \mathcal{X} \rightarrow [0, 1]$, which associate each of the 43 village configurations with a takeup rate among non-leaders.

There are no obvious a priori restrictions on the take-up rates, so we set \mathcal{F} to be all possible mappings from \mathcal{X} to $[0, 1]$. Since each $f \in \mathcal{F}$ can be identified by a vector in $[0, 1]^{43}$ and vice versa, we represent \mathcal{F} as $[0, 1]^{43}$. We set the discrepancy function as $d(f, g) := \frac{1}{43} \sum_{i=1}^{43} (f(x) - g(x))^2$ and the loss function as $l(f(x), y) := (f(x) - y)^2$.

7.2 Models

The first parametric models we consider are OLS regressions with different sets of network statistics as regressors. Specifically, we consider the following eight network statistics: (1) average eigenvector centrality of leaders; (2) average degree centrality of leaders; (3) average degree centrality of all villagers; (4) average betweenness centrality of leaders; (5) clustering coefficient of village network; (6) average path length in village network; (7) proportion of connected (non-isolated) villagers; (8) proportion of leaders.

We compute the restrictiveness and completeness of a sequence of OLS models by incrementally adding the regressors listed above. We set the baseline mapping as the OLS regression on a constant, which is a special case of all the linear models we consider. With the loss function $l(f(x), y) := (y - f(x))^2$, an estimator of completeness (computed based on in-sample errors without the use of cross validations) reduces to the R squared of the OLS regression.³⁴

³²In 2007, the microfinance institution Bharatha Swamukti Samsthe (BSE) invited a set of designated leaders within each village to an information meeting, and asked the leaders to spread the information. The data set contains the subsequent takeup rate of microfinance within each village. It additionally contains some measures of the social connections between households.

³³This is the outcome variable that Banerjee et al. (2013) focus on.

³⁴Recall that the R-squared of an OLS regression is defined by $R^2 := 1 - SSR/SST$, where $SSR := \sum_i (y_i - x_i' \hat{\beta})^2$ corresponds to the expected loss under an OLS regression model and $SST :=$

The second type of parametric model we consider is a partially linear model that we build based on the “network gossip centrality” described in Banerjee et al. (2019). To do this, we model each non-leader household’s takeup probability as a function of its position in the village.

Specifically, we define the “hearing matrix” of village i by

$$H_i(\theta_1) := \sum_{t=1}^T \theta_1^t A_i^t,$$

where T is some given number of time periods for information diffusion.³⁵ With $\theta_1 = 1$, the jk -th entry of $H_i(1)$ can be interpreted as the expected number of times villager k hears a piece of information that originates from villager j within T periods of time.³⁶ The parameter $\theta_0 \in (0, 1)$ discounts longer paths of diffusion. For each non-leader k in village i , we define

$$x_{i,k}(\theta_1) := \sum_{j \in L_i} (H_i(\theta_1))_{jk}$$

as the “network gossip centrality” of non-leader k , which counts the (discounted) sum of number of paths from the leaders of village i to non-leader k . Next, we model the takeup probability of non-leader k as function of k ’s “network gossip centrality” based on a logistic model

$$p_{i,j}(\theta_0, \theta_1) := \frac{\exp(\theta_0 + x_{i,j}(\theta_1))}{1 + \exp(\theta_0 + x_{i,j}(\theta_1))},$$

where θ_0 is a location parameter.³⁷ The (expected) village-level takeup rate among non-leaders can then be derived as the average $p_{i,j}(\theta_0, \theta_1)$ among non-leaders.

To allow additional flexibility, and to nest the naive constant model as a special case, we introduce two additional linear parameters (θ_2, θ_3) , and set:

$$f_i(\theta) := \theta_2 + \theta_3 \cdot \frac{1}{|N_i \setminus L_i|} \sum_{j \notin L_i} p_{i,j}(\theta_0, \theta_1).$$

$\sum_i (y_i - \bar{y})^2$ corresponds to the expected loss under a constant model.

³⁵We set $T = 5$, which is roughly the average diameter (of the giant components) of the 43 villages, following Banerjee et al. (2019).

³⁶ $\left(\sum_{t=1}^T A_i^t\right)_{jk}$ counts the number of paths from j to k up to length T .

³⁷Note that we do not include a scale parameter here, since if present, it will be absorbed into θ_1 .

This model is very stylized; our purpose is to illustrate how our algorithmic approach can be used to evaluate the restrictiveness of a structural model whose flexibility is otherwise difficult to gauge.

7.3 Results

Table 5: Restrictiveness and Completeness for Microfinance Takeup Rates

	# Param	Restrictiveness	Completeness
Linear Models			
Eigenvector Centrality of Leaders	1	0.9762 (0.0003)	0.2577 (0.1101)
+ Degree Centrality of Leaders	2	0.9526 (0.0004)	0.3385 (0.1193)
+ Degree Centrality of All Villagers	3	0.9288 (0.0005)	0.3471 (0.1151)
+ Betweenness Centrality of Leaders	4	0.9053 (0.0006)	0.3475 (0.1158)
+ Clustering Coefficient	5	0.8816 (0.0007)	0.3516 (0.1191)
+ Average Path Length	6	0.8579 (0.0007)	0.3516 (0.1191)
+ Proportion of Connected Villagers	7	0.8342 (0.0008)	0.3575 (0.1229)
+ Proportion of Leaders	8	0.8101 (0.0008)	0.3604 (0.1237)
Partially Linear Model	4	0.9408 (0.0036)	0.0674 (0.0452)

We report in Table 5 the restrictiveness and completeness of the models described above.³⁸ The upper panel contains results about a sequence of linear models, with a new regressor added to the OLS regression in each row.³⁹ For example, the row “Eigenvector Centrality” corresponds to an OLS regression of uptake rates on a constant and the average eigenvector centrality of leaders, while the row “+ Degree Cen-

³⁸In Table 5, the restrictiveness of the linear models is computed using $M = 10000$ simulations, while restrictiveness for the partially linear models is computed using $M = 100$ simulations. Completeness for all models is computed based the real data with $N = 43$ villages.

³⁹We add the regressors sequentially according to the ordering above, and omit the analysis of many different orderings of the same set of regressors, since the results for the current set of regressions in Table 5 are sufficient to illustrate our main point.

trality” corresponds to an OLS regression of takeup rates on a constant, the average eigenvector centrality of leaders and the average degree centrality of leaders.

The numerical results for linear models are largely as expected: as we increase the number of regressors, the model becomes increasingly flexible, so restrictiveness decreases while completeness increases. While restrictiveness seems to be decreasing at an approximately linear rate starting from the second regression, the corresponding increases in completeness appear less uniform, and in particular, completeness barely changes when we add the regressor “average path length in the village.” Note that this does not mean that this additional regressor approximately lies in the linear span of all previously included regressors, since we do observe a nontrivial reduction in restrictiveness from the addition of this regressor.

A priori it is unclear how restrictive the partially linear model is. It turns out that its restrictiveness is very high, 0.94, suggesting that the individual-by-individual modeling of takeup probabilities as a function of individual network gossip centrality imposes substantial restrictions across village configurations. However, this model’s completeness is only 0.07, so it does not capture much of the variation in village takeup rate.⁴⁰ This four-parameter partially linear model is dominated by the simple OLS model with a constant and the average eigenvector centrality of leaders as regressors: the latter has both higher restrictiveness ($0.9762 > 0.9408$) and higher completeness ($0.2577 > 0.0674$). This result shows that even a detailed, structured, and economically-motivated model may turn out to be more flexible than a simple linear model, and that this additional flexibility need not help it to fit real data.

8 Conclusion

When a theory fits the data well, it matters whether this is because the theory captures important regularities in the data, or whether the theory is so flexible that it can explain any behavior at all. We provide a practical, algorithmic approach for evaluating the restrictiveness of a theory, and demonstrate that it reveals new insights into models from two economic domains. The method is easily applied to models across diverse domains.

⁴⁰An alternative model with the two linear parameters shut down (i.e., set at $\theta_2 = 0$ and $\theta_3 = 1$) yields even higher restrictiveness and even lower completeness (that is almost 0).

In the current era of highly flexible and predictive machine learning methods, economic theory is distinguished in part by the structure it imposes on behaviors. We view these restrictions as an important part of the unique value added by economic theory. A question then naturally emerges of exactly how restrictive our models actually are compared to the nearly nonparametric approaches used in high-dimensional statistical modeling. The proposed measure offers a way to quantify this.

References

- ANDREWS, I. AND M. KASY (2019): “Identification of and Correction for Publication Bias,” *American Economic Review*, 109, 2766–2794.
- AUSTERN, M. AND W. ZHOU (2020): “Asymptotics of Cross-Validation,” *arXiv preprint arXiv:2001.11111*.
- BANERJEE, A., A. G. CHANDRASEKHAR, E. DUFLO, AND M. O. JACKSON (2013): “The diffusion of microfinance,” *Science*, 341.
- (2019): “Using gossips to spread information: Theory and evidence from two randomized controlled trials,” *The Review of Economic Studies*, 86, 2453–2490.
- BANERJEE, A., S. CHASSANG, S. MONTERO, AND E. SNOWBERG (2020): “A Theory of Experimenters: Robustness, Randomization, and Balance,” *American Economic Review*, 110, 1206–30.
- BARSEGHYAN, L., F. MOLINARI, T. O’DONOGHUE, AND J. C. TEITELBAUM (2013): “The Nature of Risk Preferences: Evidence from Insurance Choices,” *American Economic Review*, 103, 2499–2529.
- BASU, P. AND F. ECHENIQUE (2020): “On the falsifiability and learnability of decision theories,” *Theoretical Economics*, forthcoming.
- BEATTY, T. AND I. CRAWFORD (2011): “How Demanding Is the Revealed Preference Approach to Demand?” *American Economic Review*, 101, 2782–95.
- BERNHEIM, D. AND C. SPRENGER (2020): “Direct Tests of Cumulative Prospect Theory,” Working Paper.
- BRONARS, S. (1987): “The Power of Nonparametric Tests of Preference Maximization,” *Econometrica*, 55, 693–698.

- BRUHIN, A., H. FEHR-DUDA, AND T. EPPER (2010): “Risk and Rationality: Uncovering Heterogeneity in Probability Distortion,” *Econometrica*, 78, 1375–1412.
- CAMERER, C. F., T.-H. HO, AND J.-K. CHONG (2004): “A cognitive hierarchy model of games,” *The Quarterly Journal of Economics*, 119, 861–898.
- CHASSANG, S., G. P. I MIQUEL, AND E. SNOWBERG (2012): “Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments,” *American Economic Review*, 102, 1279–1309.
- CHEMLA, G. AND C. HENNESSY (2019): “Controls, belief updating, and bias in medical RCTs,” *Journal of Economic Theory*, 184.
- CHEN, X. AND A. SANTOS (2018): “Overidentification in regular models,” *Econometrica*, 86, 1771–1817.
- CHOI, S., R. FISMAN, D. GALE, AND S. KARIV (2007): “Consistency and Heterogeneity of Individual Behavior under Uncertainty,” *American Economic Review*, 97, 1–15.
- COSTA-GOMES, M., V. P. CRAWFORD, AND B. BROSETA (2001): “Cognition and behavior in normal-form games: An experimental study,” *Econometrica*, 69, 1193–1235.
- COX, D. R. (1961): “Tests of separate families of hypotheses,” in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, vol. 1, 105–123.
- (1962): “Further results on tests of separate families of hypotheses,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 406–424.
- DE QUIDT, J., J. HAUSHOFER, AND C. ROTH (2018): “Measuring and Bounding Experimenter Demand,” *American Economic Review*, 108, 3266–3302.
- DELLAVIGNA, S. AND E. LINOS (2020): “RCTs to Scale: Comprehensive Evidence from Two Nudge Units,” .
- DELLAVIGNA, S. AND D. POPE (2019): “Stability of Experimental Results: Forecasts and Evidence,” .
- FEHR-DUDA, H. AND T. EPPER (2012): “Probability and Risk: Foundations and Economic Implication of Probability-Dependent Risk Preferences,” *Annual Review of Economics*, 4, 567–593.
- FRANKEL, A. AND E. KAMENICA (2019): “Quantifying information and uncer-

- tainty,” *American Economic Review*, 109, 3650–80.
- FRANKEL, A. AND M. KASY (2019): “Which findings should be published?” Working Paper.
- FUDENBERG, D., J. KLEINBERG, A. LIANG, AND S. MULLAINATHAN (2021): “Measuring the Completeness of Theories,” Forthcoming in the *Journal of Political Economy*.
- FUDENBERG, D. AND D. LEVINE (2020): “Learning in Games and the Interpretation of Natural Experiments,” *American Economic Journal: Microeconomics*.
- FUDENBERG, D. AND A. LIANG (2019): “Predicting and Understanding Initial Play,” *American Economic Review*, 109, 4112–4141.
- GOLDSTEIN, W. M. AND H. J. EINHORN (1987): “Expression theory and the preference reversal phenomena,” *Psychological review*, 94, 236–254.
- GUL, F. (1991): “A Theory of Disappointment Aversion,” *Econometrica*, 59, 667–686.
- HANSEN, L. P. (1982): “Large sample properties of generalized method of moments estimators,” *Econometrica*, 50, 1029–1054.
- HARLESS, D. AND C. CAMERER (1994): “The Predictive Utility of Generalized Expected Utility Theories,” *Econometrica*, 62, 1251–1289.
- HAUSMAN, J. A. (1978): “Specification tests in econometrics,” *Econometrica*, 46, 1251–1271.
- HEY, J. D. (1998): “An application of Selten’s measure of predictive success,” *Mathematical Social Sciences*, 35, 1–15.
- KARMAKAR, U. (1978): “Subjectively weighted utility: A descriptive extension of the expected utility model,” *Organizational Behavior & Human Performance*, 21, 67–72.
- KOOPMANS, T. AND O. REIERSOL (1950): “The Identification of Structural Characteristics,” *The Annals of Mathematical Statistics*, 21, 165–181.
- LATTIMORE, P. K., J. R. BAKER, AND A. D. WITTE (1992): “The influence of probability on risky choice: A parametric examination,” *Journal of Economic Behavior & Organization*, 17, 315–436.
- NAGEL, R. (1995): “Unraveling in Guessing Games: An Experimental Study,” *American Economic Review*, 85, 1313–1326.

- NEWKEY, W. K. (1994): “The asymptotic variance of semiparametric estimators,” *Econometrica*, 1349–1382.
- NEWKEY, W. K. AND D. MCFADDEN (1994): “Large sample estimation and hypothesis testing,” *Handbook of Econometrics*, 4, 2111–2245.
- PEYSAKHOVICH, A. AND J. NAECKER (2017): “Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity,” *Journal of Economic Behavior and Organization*, 133, 373–384.
- POLISSON, M., J. K.-H. QUAH, AND L. RENO (2020): “Revealed Preferences over Risk and Uncertainty,” *American Economic Review*, 110, 1782–1820.
- ROUTLEDGE, B. R. AND S. E. ZIN (2010): “Generalized disappointment aversion and asset prices,” *The Journal of Finance*, 65, 1303–1332.
- SARGAN, J. D. (1958): “The estimation of economic relationships using instrumental variables,” *Econometrica*, 26, 393–415.
- SELTEN, R. (1991): “Properties for a Measure of Predictive Success,” *Mathematical Social Sciences*, 21, 153–167.
- SHMAYA, E. AND L. YARIV (2016): “Experiments on Decisions under Uncertainty: A Theoretical Framework,” *American Economic Review*, 106, 1775–1801.
- SNOWBERG, E. AND J. WOLFERS (2010): “Explaining the Favorite-Long Shot Bias: Is It Risk-Love or Misperceptions?” *Journal of Political Economy*, 118, 723–746.
- STAHL, D. O. AND P. W. WILSON (1994): “Experimental evidence on players’ models of other players,” *Journal of Economic Behavior and Organization*, 25, 309–327.
- (1995): “On players’ models of other players: Theory and experimental evidence,” *Games and Economic Behavior*, 10, 218–254.
- TVERSKY, A. AND D. KAHNEMAN (1992): “Advances in Prospect Theory: Cumulative Representation of Uncertainty,” *Journal of Risk and Uncertainty*, 5, 297–323.
- VAN DER VAART, A. W. AND J. A. WELLNER (1996): “Weak convergence,” in *Weak Convergence and Empirical Processes*, Springer, 16–28.
- VARIAN, H. (1982): “The Nonparametric Approach to Demand Analysis,” *Econometrica*, 50, 945–973.
- WRIGHT, J. R. AND K. LEYTON-BROWN (2014): “Level-0 meta-models for predicting human behavior in games,” *Proceedings of the fifteenth ACM conference on*

A Proof of Proposition 1

It is clear that A1-A4 are satisfied by the representation in (2), and A1-A5 are satisfied by the approximation error measure given in (3). For the other direction, we begin by demonstrating the following lemma:

Lemma A.1. *Suppose e satisfies A1 and A4. Then for every \mathcal{G} and d , there exists a function $h : \mathcal{F} \rightarrow \mathbb{R}$ such that*

$$e(\mathcal{G}, \mathcal{F}, d) = \mathbb{E}[h(f) : f \sim \mu_{\mathcal{F}}] \quad \forall \text{measurable } \mathcal{F}$$

Proof. Fix an arbitrary \mathcal{G} and d , and define $e_* : \Sigma \rightarrow \mathbb{R}$ to satisfy $e_*(\mathcal{F}) \equiv e(\mathcal{G}, \mathcal{F}, d)$ for all measurable \mathcal{F} . The lemma follows if we can show that A4 implies the existence of a function $h : \mathcal{F}^* \rightarrow \mathbb{R}$ such that

$$e_*(\mathcal{F}) = \int_{\mathcal{F}} h(f) d\mu_{\mathcal{F}} \quad \forall \text{measurable } \mathcal{F},$$

where $\mu_{\mathcal{F}}$ denotes the measure μ conditional on the event \mathcal{F} .

Define $\nu : \Sigma \rightarrow \mathbb{R}$ to satisfy $\nu(\mathcal{F}) = \mu(\mathcal{F}) \cdot e_*(\mathcal{F})$ for all measurable \mathcal{F} . Then A4 implies that for any countable sequence $\mathcal{F}_1, \mathcal{F}_2, \dots$,

$$\sum_{i=1}^{\infty} \nu(\mathcal{F}_i) = \nu\left(\bigcup_{i=1}^{\infty} \mathcal{F}_i\right).$$

Also, $\nu(\emptyset) = 0$ (since $\mu(\emptyset) = 0$) and ν is non-negative (by A1), so ν is a measure on (\mathcal{F}^*, Σ) . Moreover, ν is absolutely continuous with respect to μ by construction. So the Radon-Nikodym theorem implies existence of a function $h : \mathcal{F}^* \rightarrow \mathbb{R}$ such that $\nu(\mathcal{F}) = \int_{\mathcal{F}} h(f) d\mu$ for all measurable \mathcal{F} . Then

$$\mu(\mathcal{F})e_*(\mathcal{F}) = \mu(\mathcal{F}) \int_{\mathcal{F}} h(f) \frac{d\mu}{\mu(\mathcal{F})} = \mu(\mathcal{F}) \int_{\mathcal{F}} h(f) d\mu_{\mathcal{F}}$$

so $e_*(\mathcal{F}) = \int_{\mathcal{F}} h(f) d\mu_{\mathcal{F}}$, and h is exactly the function we sought. \square

Now fix any \mathcal{G} and d , and let $h_{\mathcal{G},d}$ be the function given in Lemma A.1. We will show that A2 and A3 imply that for each $f \in \mathcal{F}^*$,

$$h_{\mathcal{G},d}(f) = c_f \cdot d(\mathcal{G}, f) \tag{A.1}$$

for some constant $c_f \in \mathbb{R}_+$.

Fix an arbitrary f . Lemma A.1 implies

$$e(\mathcal{G}, \{f\}, d) = \int h_{\mathcal{G},d}(f') \cdot d\delta_f = h_{\mathcal{G},d}(f),$$

where δ_f denotes the Dirac measure at f . So it is sufficient for (A.1) to show that there is a constant $c_f \in \mathbb{R}_+$ such that $e(\mathcal{G}, \{f\}, d) = c_f \cdot d(\mathcal{G}, f)$ for all \mathcal{G}, d . By A2, models can be completely ordered for the admissible set $\{f\}$, where $e(\mathcal{G}_1, \{f\}, d) > e(\mathcal{G}_2, \{f\}, d)$ if and only if $d(\mathcal{G}_1, f) > d(\mathcal{G}_2, f)$. So there is a monotone increasing function $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ such that

$$e(\mathcal{G}, \{f\}, d) = \Phi(d(\mathcal{G}, f)). \tag{A.2}$$

Now we will show that Φ must be linear. Choose an arbitrary $\alpha \in \mathbb{R}_+$. Define $d' = \alpha \cdot d$ and suppose some model \mathcal{G}' satisfies $d(\mathcal{G}', f) = \alpha \cdot d(\mathcal{G}, f)$. Then

$$\begin{aligned} e(\mathcal{G}, \{f\}, d') &= \alpha \cdot e(\mathcal{G}, \{f\}, d) \text{ by (A3)} \\ &= \alpha \cdot \Phi(d(\mathcal{G}, f)) \text{ by (A.2)} \end{aligned}$$

and

$$\begin{aligned} e(\mathcal{G}', \{f\}, d) &= \Phi(d(\mathcal{G}', f)) \text{ by (A.2)} \\ &= \Phi(\alpha \cdot d(\mathcal{G}, f)). \end{aligned}$$

A3 requires $e(\mathcal{G}', \{f\}, d) = e(\mathcal{G}, \{f\}, d')$, so $\alpha \cdot \Phi(d(\mathcal{G}, f)) = \Phi(\alpha \cdot d(\mathcal{G}, f))$, and we have the desired linearity. Thus we can write $e(\mathcal{G}, \{f\}, d) = c_f \cdot d(\mathcal{G}, f)$ for some constant $c_f \in \mathbb{R}_+$. Repeating this argument for every f , there is a function $c : \mathcal{F} \rightarrow \mathbb{R}$ such that

$$e(\mathcal{G}, \mathcal{F}, d) = \mathbb{E}[c(f) \cdot d(\mathcal{G}, f) : f \sim \mu_{\mathcal{F}}] \quad \forall \text{measurable } \mathcal{F},$$

so we have the desired representation in (2).

Now suppose that A5 is satisfied in addition to the other axioms. The previous arguments imply that there is a function $c : \mathcal{F}^* \rightarrow \mathbb{R}$ such that

$$e(\mathcal{G}, \mathcal{F}, d) = \mathbb{E} \left[c(f) \cdot \min_{g \in \mathcal{G}} d(g, f) : f \sim \mu_{\mathcal{F}} \right] \quad \forall \mathcal{G}, \mathcal{F}, d$$

Suppose towards contradiction that e cannot be represented by (3). Then there must exist a permissible set \mathcal{F} and two mappings $f, f' \in \mathcal{F}$ such that $c(f) \cdot \mu_{\mathcal{F}}(f) > c(f') \cdot \mu_{\mathcal{F}}(f')$. But then for any models \mathcal{G}_1 and \mathcal{G}_2 with the property that

$$d(\mathcal{G}_1, f) = d(\mathcal{G}_2, f') > d(\mathcal{G}_2, f) = d(\mathcal{G}_1, f'),$$

it follows that $e(\mathcal{G}_1, \{f, f'\}, d) > e(\mathcal{G}_2, \{f, f'\}, d)$, violating A5.

B Estimation of Completeness κ

B.1 Preliminary Definitions

We now introduce some definitions and notation that will be useful in the derivation of the asymptotic distribution of the CV-based completeness estimator.

B.1.1 Finite-Sample Out-of-Sample Error

Let $\mathbf{Z}_N := (Z_i)_{i=1}^N$ be a random sample of observations in a given data set, and let $Z_{N+1} \sim P^*$ denote a random variable with the same distribution P^* that is independent of \mathbf{Z}_N . For a given data set \mathbf{Z}_N and a given model $\tilde{\mathcal{F}}$, we define the conditional out-of-sample error (given data set \mathbf{Z}_N) as

$$e_{\tilde{\mathcal{F}}}(\mathbf{Z}_N) := \mathbb{E} \left[l \left(\hat{f}_{\mathbf{Z}_N}, Z_{N+1} \right) \middle| \mathbf{Z}_N \right],$$

where $\hat{f}_{\mathbf{Z}_N} \in \tilde{\mathcal{F}}$ is an estimator, or an algorithm, that selects a mapping $\hat{f}_{\mathbf{Z}_N}$ within the model $\tilde{\mathcal{F}}$ based on data \mathbf{Z}_N . We also define the out-of-sample error, with expectation taken over different possible data sets \mathbf{Z}_N , as $e_{\tilde{\mathcal{F}}, N} := \mathbb{E} [e_{\tilde{\mathcal{F}}}(\mathbf{Z}_N)]$.

From the definition of the K-fold cross-validation estimator, it can be easily

shown that $\mathbb{E} \left[\hat{e}_{CV} \left(\tilde{\mathcal{F}} \right) \right] = e_{\mathcal{F}, \frac{K-1}{K}N}$. As a result, the asymptotic distribution of $\hat{e}_{CV} \left(\tilde{\mathcal{F}} \right) - e_{\mathcal{F}, \frac{K-1}{K}N}$ has been studied in the statistics and machine learning literature. Our analysis below will be based on the results in Austern and Zhou (2020) on the asymptotic distribution of $\hat{e}_{CV} \left(\tilde{\mathcal{F}} \right) - e_{\mathcal{F}, \frac{K-1}{K}N}$.

B.1.2 Joint Parametrization of \mathcal{G} and \mathcal{F}

Recall that the model \mathcal{G} is parametrized by $\theta \in \Theta$, and f_θ denotes a generic function in \mathcal{G} . Motivated by the applications in this paper, we assume that \mathcal{F} can be smoothly parameterized by a finite-dimensional parameter $\beta \in \mathcal{B} \subseteq \mathbb{R}^{d_{\mathcal{F}}}$ and use the notation $f_{[\beta]} \in \mathcal{F}$ to denote a generic function in \mathcal{F} . Since by assumption $f^* \in \mathcal{F}$, we can define a parameter β^* to represent it, i.e. $f_{[\beta^*]} = f^*$.

For arbitrary parameters θ and β , write $l_\Theta(\theta, Z_i) := l(f_\theta, Z_i)$ and $l_{\mathcal{B}}(\beta, Z_i) := l(f_{[\beta]}, Z_i)$. We define the estimation mappings in \mathcal{G} and \mathcal{F} by

$$\begin{aligned} \hat{\theta}(\mathbf{Z}_N) &:= \arg \min_{\theta \in \Theta} \frac{1}{N} \sum l_\Theta(\theta, Z_i), \\ \hat{\beta}(\mathbf{Z}_N) &:= \arg \min_{\beta \in \mathcal{B}_{\mathcal{M}}} \frac{1}{N} \sum l_{\mathcal{B}}(\beta, Z_i). \end{aligned}$$

Let $\alpha := (\theta', \beta)'$ denote the concatenation of the parameters $\theta \in \Theta$ and $\beta \in \mathcal{B}$, $\alpha^* := (\theta^*, \beta^*)'$ to be the parameters associated with the best mappings in \mathcal{G} and \mathcal{F} , and also define

$$\hat{\alpha}(\mathbf{Z}_N) := \left(\hat{\theta}'(\mathbf{Z}_N), \hat{\beta}'(\mathbf{Z}_N) \right)' = \arg \min_{\theta \in \Theta, \beta \in \mathcal{B}} \frac{1}{N} \sum_{i=1}^N [l_\Theta(\theta, Z_i) + l_{\mathcal{B}}(\beta, Z_i)]$$

to be an estimator for α^* . Finally, define

$$\Delta l(\theta, \beta; Z_i) := l(f_\theta, Z_i) - l(f_{[\beta]}, Z_i) = l_\Theta(\theta, Z_i) - l_{\mathcal{B}}(\beta, Z_i).$$

B.2 Construction of Variance Estimator

To obtain the standard error of the estimate, we use a variance estimator adapted from Proposition 1 in Austern and Zhou (2020). Specifically, for the k -th test set, let $f_{\hat{\theta}^{-k}}$ and \hat{f}^{-k} be the estimated mappings from models \mathcal{G} and \mathcal{F} , respectively. The

difference in their test errors on observation Z_i is

$$\Delta_\theta(Z_i) := l(f_{\hat{\theta}^{-k}}, Z_i) - l(\hat{f}^{-k}, Z_i),$$

and the average difference across all observations in test fold k is

$$\bar{\Delta}_{\theta,k} := \frac{1}{J_N} \sum_{k(i)=k} \Delta_\theta(Z_i).$$

The sample variance of the difference in test errors for the k -th fold is

$$\hat{\sigma}_{\Delta_\theta,k}^2 := \frac{1}{J_N - 1} \sum_{k(i)=k} (\Delta_\theta(Z_i) - \bar{\Delta}_{\theta,k})^2$$

which we then average over the K folds and obtain

$$\hat{\sigma}_{\Delta_\theta}^2 := \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_{\Delta_\theta,k}^2.$$

Similarly we define $\Delta_{f_{\text{base}}}(Z_i) := l(f_{\text{base}}, Z_i) - l(\hat{f}^{-k}, Z_i)$, and correspondingly $\bar{\Delta}_{f_{\text{base}},k}$, $\hat{\sigma}_{\Delta_{f_{\text{base}}},k}^2$ and $\hat{\sigma}_{\Delta_{f_{\text{base}}}}^2$. Lastly, define the covariance estimator by

$$\hat{\sigma}_{\Delta_\theta \Delta_{f_{\text{base}}}} := \frac{1}{K} \sum_{k=1}^K \frac{1}{J_N - 1} \sum_{k(i)=k} (\Delta_\theta(Z_i) - \bar{\Delta}_{\theta,k}) (\Delta_{f_{\text{base}}}(Z_i) - \bar{\Delta}_{f_{\text{base}},k}(Z_i)).$$

Based on $\hat{\sigma}_{\Delta_\theta}^2$, $\hat{\sigma}_{\Delta_{f_{\text{base}}}}^2$ and $\hat{\sigma}_{\Delta_\theta \Delta_{f_{\text{base}}}}$, we define the following variance estimator for $\hat{\kappa}$:

$$\hat{\sigma}_{\hat{\kappa}}^2 := \frac{\hat{\sigma}_{\Delta_\theta}^2 - 2\hat{\kappa}\hat{\sigma}_{\Delta_\theta \Delta_{f_{\text{base}}}} + \hat{\kappa}^2\hat{\sigma}_{\Delta_{f_{\text{base}}}}^2}{[\hat{e}_{CV}(f_{\text{base}}) - \hat{e}_{CV}(\mathcal{F})]^2}. \quad (\text{B.1})$$

B.3 Assumptions and Lemmas Based on Austern and Zhou (2020)

Assumption 3 (Conditions for Asymptotics of CV Estimator).

1. $l_\Theta(\theta, z)$ and $l_B(\beta, z)$ are twice differentiable and strictly convex in θ and β .
2. $\mathbb{E}[\sup_{\theta \in \Theta} l_\Theta^4(\theta, Z_i)] < \infty$ and $\mathbb{E}[\sup_{\beta \in B} l_B^4(\beta, Z_i)] < \infty$.

3. There exist open neighborhoods \mathcal{O}_{θ^*} and \mathcal{O}_{β^*} of θ^* and β^* in Θ and \mathcal{B} such that

- (a) $\mathbb{E} \left[\sup_{\theta \in \mathcal{O}_{\theta^*}} \|\nabla_{\theta} l_{\Theta}(\theta, Z_i)\|^{16} \right] < \infty$, $\mathbb{E} \left[\sup_{\beta \in \mathcal{O}_{\beta^*}} \|\nabla_{\beta} l_{\mathcal{B}}(\beta, Z_i)\|^{16} \right] < \infty$.
- (b) $\mathbb{E} \left[\sup_{\theta \in \mathcal{O}_{\theta^*}} \|\nabla_{\theta}^2 l_{\Theta}(\theta, Z_i)\|^{16} \right] < \infty$, $\mathbb{E} \left[\sup_{\beta \in \mathcal{O}_{\beta^*}} \|\nabla_{\beta}^2 l_{\mathcal{B}}(\beta, Z_i)\|^{16} \right] < \infty$.
- (c) there exists $c > 0$ such that $\lambda_{\min}(\nabla_{\theta}^2 l_{\Theta}(\theta, Z_i)) \geq c$, $\lambda_{\min}(\nabla_{\beta}^2 l_{\mathcal{B}}(\beta, Z_i)) \geq c$ a.s. uniformly on \mathcal{O}_{θ^*} and \mathcal{O}_{β^*} .

Lemma B.1 (Application of Proposition 5 of Austern and Zhou, 2020). *Under Assumption 3:*

$$\sqrt{N} \left[\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - \left(e_{\mathcal{G}, \frac{K-1}{K}N} - e_{\mathcal{F}, \frac{K-1}{K}N} \right) \right] \xrightarrow{d} \mathcal{N} \left(0, \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i)) \right).$$

Proof. Proposition 5 of Austern and Zhou (2020) establishes the asymptotic normality of cross-validation risk estimator and its asymptotic variance under parametric settings where the loss function used for training is the same as the loss function used for evaluation. Applying Proposition 5 of Austern and Zhou (2020) under Assumption 3 to θ, β and $\alpha = (\theta, \beta)$, we obtain:

$$\begin{aligned} & \sqrt{N} \left(\hat{e}_{CV}(\mathcal{G}) - e_{\mathcal{G}, \frac{K-1}{K}N} \right) \xrightarrow{d} \mathcal{N} \left(0, \text{Var}(l(f_{\theta^*}, Z_i)) \right), \\ & \sqrt{N} \left(\hat{e}_{CV}(\mathcal{F}) - e_{\mathcal{F}, \frac{K-1}{K}N} \right) \xrightarrow{d} \mathcal{N} \left(0, \text{Var}(l(f^*, Z_i)) \right), \\ & \sqrt{N} \left(\hat{e}_{CV}(\mathcal{G}) + \hat{e}_{CV}(\mathcal{F}) - e_{\mathcal{G}, \frac{K-1}{K}N} - e_{\mathcal{F}, \frac{K-1}{K}N} \right) \xrightarrow{d} \mathcal{N} \left(0, \text{Var}(l(f_{\theta^*}, Z_i) + l(f^*, Z_i)) \right). \end{aligned}$$

Using the equality $\text{Var}(X + Y) + \text{Var}(X - Y) = 2\text{Var}(X) + 2\text{Var}(Y)$, we then deduce that

$$\sqrt{N} \left[\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - \left(e_{\mathcal{G}, \frac{K-1}{K}N} - e_{\mathcal{F}, \frac{K-1}{K}N} \right) \right] \xrightarrow{d} \mathcal{N} \left(0, \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i)) \right).$$

□

Lemma B.2 (Application of Proposition 1 of Austern and Zhou, 2020). *Under Assumption 3, $\hat{\sigma}_{\Delta}^2 \xrightarrow{p} \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i))$.*

Proof. Applying Proposition 1 of Austern and Zhou (2020) under Assumption 3 to

θ, β and $\alpha = (\theta, \beta)$:

$$\begin{aligned} \hat{\sigma}_{\mathcal{G}}^2 &:= \frac{1}{K} \sum_{k=1}^K \frac{1}{J_N - 1} \sum_{k(i)=k} \left(l(f_{\hat{\theta}^{-k}}, Z_i) - \frac{1}{J_N} \sum_{k(j)=k} l(f_{\hat{\theta}^{-k}}, Z_j) \right)^2 \\ &\xrightarrow{p} \text{Var}(l(f_{\theta^*}, Z_i)). \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}_{\mathcal{F}}^2 &:= \frac{1}{K} \sum_{k=1}^K \frac{1}{J_N - 1} \sum_{k(i)=k} \left(l(f_{[\hat{\beta}^{-k}]}, Z_i) - \frac{1}{J_N} \sum_{k(j)=k} l(f_{[\hat{\beta}^{-k}]}, Z_j) \right)^2 \\ &\xrightarrow{p} \text{Var}(l(f^*, Z_i)). \end{aligned}$$

and

$$\begin{aligned} &\hat{\sigma}_{\mathcal{G}+\mathcal{F}}^2 \\ &:= \frac{1}{K} \sum_{k=1}^K \frac{1}{J_N - 1} \cdot \sum_{k(i)=k} \left(l(f_{\hat{\theta}^{-k}}, Z_i) + l(f_{[\hat{\beta}^{-k}]}, Z_i) - \frac{1}{J_N} \sum_{k(j)=k} [l(f_{[\hat{\beta}^{-k}]}, Z_j) + l(f_{\hat{\theta}^{-k}}, Z_i)] \right)^2 \\ &\xrightarrow{p} \text{Var}(l(f_{\theta^*}, Z_i) + l(f^*, Z_i)), \end{aligned}$$

Hence:

$$\begin{aligned} \hat{\sigma}_{\Delta_\theta}^2 &= 2\hat{\sigma}_{\mathcal{G}}^2 + 2\hat{\sigma}_{\mathcal{F}}^2 - \hat{\sigma}_{\mathcal{G}+\mathcal{F}}^2 \\ &\xrightarrow{p} 2\text{Var}(l(f_{\theta^*}, Z_i)) + 2\text{Var}(l(f^*, Z_i)) - 2\text{Var}(l(f_{\theta^*}, Z_i) + l(f^*, Z_i)) \\ &= \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i)) \end{aligned}$$

□

B.4 Proof of Proposition 3

Lemma B.1 characterizes the limit distribution of

$$\sqrt{N} \left[\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - \left(e_{\mathcal{G}, \frac{K-1}{K}N} - e_{\mathcal{F}, \frac{K-1}{K}N} \right) \right]$$

which we show is also the limit distribution of $\sqrt{N} [\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - (e_{\mathcal{G}} - e_{\mathcal{F}})]$.

To see this, notice that

$$\begin{aligned} & e_{\mathcal{G}, \frac{K-1}{K}N} - e_{\mathcal{G}} \\ &= \mathbb{E} \left[l_{\Theta} \left(\hat{\theta}^{-k(i)}, Z_i \right) - l_{\Theta} \left(\theta^*, Z_i \right) \right] \\ &= \mathbb{E} \left[\nabla l_{\Theta} \left(\theta^*, Z_i \right) \cdot \left(\hat{\theta}^{-k(i)} - \theta^* \right) + \left(\hat{\theta}^{-k(i)} - \theta^* \right)' \nabla^2 l_{\Theta} \left(\tilde{\theta}, Z_i \right) \cdot \left(\hat{\theta}^{-k(i)} - \theta^* \right) \right] \\ &= 0 + \mathbb{E} \left[\left(\hat{\theta}^{-k(i)} - \theta^* \right)' \nabla^2 l_{\Theta} \left(\tilde{\theta}, Z_i \right) \cdot \left(\hat{\theta}^{-k(i)} - \theta^* \right) \right] \\ &= \frac{1}{N - J_N} \mathbb{E} \left[\sqrt{N - J_N} \left(\hat{\theta}^{-k(i)} - \theta^* \right)' \nabla^2 l_{\Theta} \left(\tilde{\theta}, Z_i \right) \cdot \sqrt{N - J_N} \left(\hat{\theta}^{-k(i)} - \theta^* \right) \right] \\ &= c \frac{1}{N - J_N} + o \left(\frac{1}{N - J_N} \right) = c \frac{K}{K - 1} \cdot \frac{1}{N} + o \left(\frac{1}{N} \right) \end{aligned}$$

since $J_N = N/K$. Therefore

$$\sqrt{N} \left(e_{\Theta, \frac{K-1}{K}N} - e_{\Theta} \right) = o_p(1).$$

and, similarly, $\sqrt{N} \left(e_{\mathcal{F}, \frac{K-1}{K}N} - e_{\mathcal{F}} \right) = o_p(1)$. Hence:

$$\sqrt{N} [\hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - (e_{\mathcal{G}} - e_{\mathcal{F}})] \xrightarrow{d} \mathcal{N}(0, \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i))).$$

Now, we replicate the previous result with f_{base} in place of \mathcal{G} and obtain

$$\sqrt{N} [\hat{e}_{CV}(f_{\text{base}}) - \hat{e}_{CV}(\mathcal{F}) - (e_{f_{\text{base}}} - e_{\mathcal{F}})] \xrightarrow{d} \mathcal{N}(0, \text{Var}(\Delta l(f_{\text{base}}, f^*; Z_i))).$$

and jointly

$$\sqrt{N} \begin{pmatrix} \hat{e}_{CV}(\mathcal{G}) - \hat{e}_{CV}(\mathcal{F}) - (e_{\mathcal{G}} - e_{\mathcal{F}}) \\ \hat{e}_{CV}(f_{\text{base}}) - \hat{e}_{CV}(\mathcal{F}) - (e_{f_{\text{base}}} - e_{\mathcal{F}}) \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \begin{pmatrix} \sigma_{\Delta_{\theta}}^2 & \sigma_{\Delta_{\theta}\Delta_{f_{\text{base}}}} \\ \sigma_{\Delta_{\theta}\Delta_{f_{\text{base}}}} & \sigma_{\Delta_{f_{\text{base}}}}^2 \end{pmatrix} \right)$$

with

$$\begin{aligned} \sigma_{\Delta_{\theta}}^2 &:= \text{Var}(\Delta l(f_{\theta^*}, f^*; Z_i)) \\ \sigma_{\Delta_{f_{\text{base}}}}^2 &:= \text{Var}(\Delta l(f_{\text{base}}, f^*; Z_i)) \\ \sigma_{\Delta_{\theta}\Delta_{f_{\text{base}}}} &:= \text{Cov}(\Delta l(f_{\theta^*}, f^*; Z_i), \Delta l(f_{\text{base}}, f^*; Z_i)) \end{aligned}$$

By Lemma B.2, Assumption 2 and the Delta Method, we have

$$\sqrt{N}(\hat{\kappa} - \kappa) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_{\Delta_{\theta}}^2 - 2\kappa\sigma_{\Delta_{\theta}\Delta_{f_{\text{base}}}} + \kappa^{*2}\sigma_{\Delta_{f_{\text{base}}}}^2}{d^2(f_{\text{base}}, f^*)} \right)$$

Since

$$\hat{\sigma}_{\hat{\kappa}} \xrightarrow{p} \frac{\sigma_{\Delta_{\theta}}^2 - 2\kappa\sigma_{\Delta_{\theta}\Delta_{f_{\text{base}}}} + \kappa^{*2}\sigma_{\Delta_{f_{\text{base}}}}^2}{d^2(f_{\text{base}}, f^*)},$$

we have

$$\frac{\sqrt{N}(\hat{\kappa} - \kappa)}{\hat{\sigma}_{\hat{\kappa}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

C Estimates for Application 1

Table 6: Restrictiveness and Completeness for Certainty Equivalents

	# Param	Restrictiveness	Completeness
CPT Specifications			
α, δ, γ	3	0.28 (0.003)	0.95 (0.02)
δ, γ	2	0.37 (0.004)	0.95 (0.02)
α, γ	2	0.51 (0.006)	0.95 (0.02)
α, δ	2	0.49 (0.005)	0.27 (0.05)
α	1	0.91 (0.005)	0.25 (0.05)
δ	1	0.68 (0.009)	0.26 (0.06)
γ	1	0.59 (0.006)	0.71 (0.06)
DA Specifications			
α, η	2	0.47 (0.006)	0.27 (0.06)
η	1	0.69 (0.009)	0.27 (0.05)

Restrictiveness is estimated from 1000 simulations and we report their analytic standard errors. Because of potential dependence among the reported certainty equivalents of subjects, we compute the standard errors for completeness using a block bootstrapping procedure that clusters together all observations from the same subject.⁴¹ We then carry out our (cross-validated) estimation of completeness on each bootstrap sample, and compute the standard errors based on 1000 bootstrap samples. These bootstrapped standard errors are similar to the analytic standard errors we get under a revision of the formulas in Section 4 to accommodate clustering on subjects (see Online Appendix O.6).

⁴¹Specifically, when generating a bootstrap sample, we randomly sample from the list of 179 subjects with replacement, and include all the reported certainty equivalents of the drawn subjects with replacement.

O For Online Publication

O.1 Parameter Estimates

We report below the estimated parameters for each of the models that we consider. In the first column, we report the estimated parameters on the actual data. In the second, we report the average parameter estimates across our generated mappings.

CPT Parameters	Real Data	Synthetic Data
α, δ, γ	(0.78,0.98,0.42)	(0.65,3.95,0.43)
α, γ	(0.77,0.40)	(0.91,0.27)
α, δ	(1,0.68)	(0.43,17.74)
δ, γ	(0.70,0.50)	(1.80,0.42)
α	0.77	0.87
γ	0.44	0.25
δ	0.68	5.15

DA Parameters	Real Data	Synthetic Data
α, η	(1,0.47)	(0.35,-0.77)
η	0.47	-0.20

	Real Data	Synthetic Data
PCHM	$\tau = 0.5$	$\tau = 0.1$
logit level-1	$\lambda = 0.02$	$\lambda = 0.0018$
logit PCHM	$(\tau, \lambda) = (1.4, 0.11)$	$(\tau, \lambda) = (1.05, 0.02)$

O.2 Heterogeneous Risk Preferences

Our analysis in the main text considered representative agent models. In some cases, the analyst may have auxiliary data on the subjects that can be used to improve predictions. We show now how completeness and restrictiveness can be evaluated in this case.

Specifically, we return to our first application and group subjects into three clusters identified by Bruhin et al. (2010). We fit CPT for each cluster, allowing parameter values to vary across groups. Table O.2 reports completeness measures cluster by cluster.

	Cluster 1	Cluster 2	Cluster 3
Baseline Mapping	39.90 (4.98)	150.10 (7.24)	99.94 (7.97)
CPT	30.74 (7.25)	43.87 (4.72)	69.62 (8.50)
Best Achievable Error	29.59 (7.36)	36.30 (3.34)	67.05 (8.02)
Completeness	0.98 (0.02)	0.88 (0.03)	0.92 (0.03)
N	674	1144	2641

Table 7: Completeness measures for each of three subject clusters.

The performance of the expected value rule, the best achievable performance, and the performance of CPT, all vary substantially across clusters. For example, the behavior of subjects in cluster 1 is roughly consistent with expected value (the error of the expected value rule is 39.90), while the behavior of subjects in cluster 3 departs substantially from this benchmark (the error of the expected value rule is 99.94). The best achievable prediction for these groups of subjects is also very different (ranging from 29.59 to 67.05), as is the completeness of CPT (ranging from 30.74 to 69.62).

The average completeness, weighted by proportion of observations in each cluster, is 0.91, which is very close to what we found for the representative agent model. This may seem surprising at first, since allowing for parameters to vary across subjects improves the accuracy of predictions. But the best mapping from the extended feature space $\mathcal{X}' = \mathcal{X} \times \{1, 2, 3\}$ to \mathcal{Y} is more predictive than the best mapping considered previously. Thus what we find is that the completeness of CPT with three clusters, relative to the best three-cluster mapping, is comparable to the completeness of the representative-agent version of CPT, relative to the best representative-agent mapping.

Similarly, when measuring restrictiveness, we extend the set of admissible mappings to the domain \mathcal{X}' . Each generated pattern of behavior is thus a triple (f_1, f_2, f_3)

of mappings from the original \mathcal{F} . We ask how well these tuples can be approximated using mappings (g_1, g_2, g_3) from CPT. It is straightforward to see that the restrictiveness of the three-cluster CPT is identical to the restrictiveness of the representative-agent model.⁴²

O.3 Relationship between Completeness and Restrictiveness

In this section, we show that completeness and restrictiveness are related via the equation

$$\kappa(\mathcal{G}) = 1 - r(\mathcal{G}, \{f^*\}), \quad (\text{O.1})$$

when the loss function l used to define e_{P^*} , and the discrepancy function d used to define r , are “paired” in a coherent way, which we now explain.

We first provide more details about the formulation of completeness. Suppose that besides X , there is a random outcome Z . We will consider hypothetical joint distributions P that share a common marginal distribution P_X^* . Fixing any such distribution P , the analyst wants to learn a statistic of the conditional distribution of Z given X , which we denote by $y(P_{Z|X=x}) \in \mathcal{Y}$. Two leading cases of this problem are: (a) prediction of the conditional expectation $\mathbb{E}_P[Z|X]$, and (b) prediction of the conditional distribution $P_{Z|X}$ itself. As in the main text, a mapping is any function $f : \mathcal{X} \rightarrow \mathcal{Y}$, and we define \mathcal{F}^* to be the set of all such mappings.

Let $l : \mathcal{F}^* \times \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function, where $l(f, (x, z))$ is the loss assigned to predicting $f(x)$ when the realized outcome is z . When $(X, Z) \sim P$, the expected error of mapping f is

$$e_P(f) := \mathbb{E}_P[l(f, (X, Z))] \quad (\text{O.2})$$

and the mapping that minimizes expected error is

$$f_P = \min_{f \in \mathcal{F}} e_P(f).$$

If we let P^* denote the joint distribution from which real data is drawn, then the

⁴²Note that this is true for any number of exogenously specified clusters.

completeness of a model \mathcal{G} as defined in Fudenberg et al. (2021) is

$$\kappa(\mathcal{G}) = \frac{e_{P^*}(f_{\text{base}}) - e_{P^*}(\mathcal{G})}{e_{P^*}(f_{\text{base}}) - e_{P^*}(\mathcal{F}^*)} \equiv 1 - \frac{e_{P^*}(\mathcal{G}) - e_{P^*}(\mathcal{F}^*)}{e_{P^*}(f_{\text{base}}) - e_{P^*}(\mathcal{F}^*)}.$$

We now formally define the meaning of “pairing” between the discrepancy function d and the loss function l .

Definition O.1. The loss function l and discrepancy $d : \mathcal{F}^* \times \mathcal{F}^* \rightarrow \mathbb{R}$ are *paired* if

$$d(f, f_P) = e_P(f) - e_P(f_P) \tag{O.3}$$

for every distribution $P \in \Delta(\mathcal{X} \times \mathcal{Z})$ whose marginal distribution on \mathcal{X} is P_X^* . That is, $d(f, f_P)$ is the difference between the error of mapping f and the error of the best mapping f_P .⁴³

As noted in the main text, if l and d are paired, then (O.1) holds, where $f^* = f_{P^*}$. Moreover, as also noted in the main text, the following functions are paired:

- Let $\mathcal{Y} = \mathbb{R}$. Then squared loss $l(f, (x, z)) := (z - f(x))^2$ and the squared distance discrepancy $d_{MSE}(f, g) := \mathbb{E}_{P_X^*} [(f(X) - g(X))^2]$ are paired.
- Let \mathcal{Y} be the set of distributions over a finite set \mathcal{Z} . Then negative (conditional) log-likelihood $l(f, (x, z)) := -\log f(z|x)$ and the KL-divergence discrepancy

$$d_{KL}(f, g) := \mathbb{E}_{P_X^*} \left[\sum_{z \in \mathcal{Z}} g(z|x) [\log g(z|x) - \log f(z|x)] \right]$$

are paired.

O.3.1 A Loss Function That Cannot be Paired with any Discrepancy

When \mathcal{Y} is the set of distributions on \mathcal{Z} , then every loss function l has a paired discrepancy function, since we can define $d(f, f_P) := e_{f_P}(f) - e_{f_P}(f_P)$.⁴⁴ But in

⁴³This relation resembles but differs from the coupling of the “cost of uncertainty” and the “value of information” in Frankel and Kamenica (2019), which concerns comparisons of different signal structures, as opposed to comparing model classes.

⁴⁴This is because P is completely pinned down by f_P given P_X^* , so $e_P = e_{f_P}$.

general, for some prediction problems and loss functions l , there may not exist a discrepancy d such that l and d are paired, as the next example shows. In these cases, we can still evaluate restrictiveness and completeness, but they will not have an evident relationship.

Consider a setting where X is degenerate, i.e., \mathcal{X} is a singleton, so that the joint distribution P is completely characterized by the distribution of Y . Furthermore, let $\mathcal{Y} := [0, 1]$.

If $f^* := \text{med}(Y) \in \mathcal{Y} = [0, 1]$, then a mapping $f : \mathcal{X} \rightarrow \mathcal{S}$ is just a number in $[0, 1]$. When the loss function is the absolute deviation $l(f, y) := |y - f|$, and the error function is mean absolute deviation $e_{P^*}(f) := \mathbb{E}_{P^*}[|Y - f|]$, the true median f^* minimizes the error, i.e. $f^* \in \arg \min_{f \in [0, 1]} e_{P^*}(f)$. However, it is not true that $|f - f^*| = e_{P^*}(f) - e_{P^*}(f^*)$ for any $f \in [0, 1]$. To see this, suppose that $Y \sim U[0, 1]$ under P^* . Then $f^* = 0.5$ and $e_{P^*}(f^*) = 0.25$. However, for $f = 0.4$, we have $e_{P^*}(f) = 0.26$. but $|f - f^*| = 0.1 \neq 0.01 = e_{P^*}(f) - e_{P^*}(f^*)$.

Moreover, there is no function $d : [0, 1]^2 \rightarrow [0, 1]$ such that decomposability (O.3) holds, which would require that $d(f, f_P) = e_P(f) - e_P(f_P)$ for any distribution P of Y supported on $[0, 1]$. To see this, suppose that $Y \sim U[0, 1]$ under P_1 , we have

$$e_{P_1}(f) - e_{P_1}(f_{P_1}) = (f - 0.5)^2 = (f - f_{P_1})^2, \quad \forall f \in [0, 1].$$

However, supposing that, under P_2 , the probability density function of Y is given by $2y$ for $y \in [0, 1]$, we have $f_{P_2} = \sqrt{2}/2$ and $e_{P_2}(f_{P_2}) = (2 - \sqrt{2})/3$ but

$$e_{P_2}(f) - e_{P_2}(f_{P_2}) = \frac{1}{3} \left(2f^3 - 3f^2 + \sqrt{2} \right) \neq (f - f_{P_2})^2.$$

O.4 Extension to Infinite-Dimensional \mathcal{F}

Now we consider a setting where \mathcal{X} , the support of X , is a continuum, and the set of admissible mappings \mathcal{F} is an infinite-dimensional function space. For simplicity, we focus on the case of a compact and rectangular $\mathcal{X} := [0, 1]^{d_x}$.

O.4.1 Computing Restrictiveness r

Here the distribution μ on \mathcal{F} is a distribution over an infinite-dimensional functional space, so “direct” simulation from μ becomes infeasible. We propose two ways to proceed:

Simulation on a Growing Grid

For a given number of simulations M , we can restrict our attention to a finite grid of the form

$$\mathcal{X}_M := \left\{ \frac{k}{K_M} : k = 0, 1, \dots, 2^{K_M} \right\}^{d_x},$$

where $K_M \rightarrow \infty$ and $K_M^{d_x}/M \rightarrow 0$ as $M \rightarrow \infty$. We then proceed by simulating f from a measure (say, uniform) μ_M on the restriction of \mathcal{F} on \mathcal{X}_M .

Simulation of Coefficients on Basis Functions

Alternatively, if \mathcal{F} satisfies certain smoothness conditions (e.g. possess uniformly bounded derivatives up to a certain order), then we can specify a sequence of orthonormal basis functions $\{b_k(x) : k \in \mathbb{N}\}$, such as (tensor products of) power series, trigonometric series, splines, wavelets (See, e.g., Chen 2007 for a survey), such that the linear span of the basis functions is dense in \mathcal{F} . Shape restrictions in \mathcal{F} , such as nonnegativity, monotonicity and convexity, can also be incorporated by proper specification of the basis functions. Writing

$$\mathcal{B}_M := \left\{ f_{[\beta]} := \sum_{k=1}^{K_M} \beta_k b_k(x) : \beta \in \mathbb{R}^{K_M} \right\}$$

for some $K_M \rightarrow \infty$ as $M \rightarrow \infty$, we could specify simulate f from \mathcal{B}_M by randomly drawing β from some measure $\mu_{\beta, M}$ on \mathbb{R}^{K_M} .

O.4.2 Estimating Completeness κ

The estimation of completeness κ can be also adapted to accommodate an infinite-dimensional \mathcal{F} , either via the growing-grid approach or the basis-function approach.

We illustrate the asymptotic property of $\hat{\kappa}$ using the later approach, and focus on a simpler setting without the use of cross validations.

Specifically, for a given loss function l and a model $\mathcal{G} := f_\theta : \{\theta \in \Theta\}$ parameterized by θ , define

$$\begin{aligned}\hat{e}(f) &:= \frac{1}{N} \sum_{i=1}^N l(Z_i, f) \\ \hat{\theta} &:= \arg \min_{\theta \in \Theta} \hat{e}(f_\theta) \\ \hat{\beta} &:= \arg \min_{\beta \in \mathcal{B}_{M_N}} \hat{e}(f_{[\beta]}) \\ \hat{e}(\mathcal{G}) &:= \hat{e}(f_{\hat{\theta}}) \\ \hat{e}(\mathcal{F}) &:= \hat{e}(f_{[\hat{\beta}]}) .\end{aligned}$$

with $M_N \rightarrow \infty$ slowly and

$$\hat{\kappa} := 1 - \frac{\hat{e}(\mathcal{G}) - \hat{e}(\mathcal{F})}{\hat{e}(f_{\text{base}}) - \hat{e}(\mathcal{F})} .$$

Under standard regularity conditions for nonparametric estimation,

$$\|f_{\hat{\beta}} - f^*\| \xrightarrow{p} 0, \text{ as } N \rightarrow \infty .$$

Observing

$$\frac{1}{N} \sum_{i=1}^N [l(Z_i, f_{\hat{\theta}}) - \hat{e}(\mathcal{G})] = 0, \quad \frac{1}{N} \sum_{i=1}^N \nabla_{\theta} l(Z_i, f_{\hat{\theta}}) = 0,$$

and, by the definition of θ^* ,

$$\mathbb{E}[\nabla_{\theta} l(Z_i, f_{\theta^*})] = 0$$

we have, by Theorem 6.1 of Newey and McFadden (1994),

$$\sqrt{N} [\hat{e}(\mathcal{G}) - e(\mathcal{G})] \xrightarrow{d} \mathcal{N}(0, \text{Var}[l(Z_i, f_{\theta^*})]) .$$

Similarly, since

$$\frac{1}{N} \sum_{i=1}^N \left[l \left(Z_i, f_{[\hat{\beta}]} \right) - \hat{e}(\mathcal{F}) \right] = 0, \quad \frac{1}{N} \sum_{i=1}^N \nabla_{\beta} l \left(Z_i, f_{[\hat{\beta}]} \right) = 0,$$

and

$$\mathbb{E} \left[\nabla_{\beta} l \left(Z_i, f_{[\beta^*]} \right) \right] = 0,$$

we have, by Proposition 2 of Newey (1994),

$$\sqrt{N} \left[\hat{e}(\mathcal{F}) - e(\mathcal{F}) \right] \xrightarrow{d} \mathcal{N} \left(0, \text{Var} \left[l \left(Z_i, f^* \right) \right] \right).$$

It is then straightforward to extend the above to obtain:

$$\sqrt{N} \begin{pmatrix} \hat{e}(\mathcal{G}) - \hat{e}(\mathcal{F}) - e(\mathcal{G}) + e(\mathcal{F}) \\ \hat{e}(f_{\text{base}}) - \hat{e}(\mathcal{F}) - e(f_{\text{base}}) + e(\mathcal{F}) \end{pmatrix} \xrightarrow{d} \mathcal{N} \left(0, \begin{pmatrix} \sigma_{\Delta_{\theta}}^2 & \sigma_{\Delta_{\theta} \Delta_{f_{\text{base}}}} \\ \sigma_{\Delta_{\theta} \Delta_{f_{\text{base}}}} & \sigma_{\Delta_{f_{\text{base}}}}^2 \end{pmatrix} \right)$$

and

$$\sqrt{N} (\hat{\kappa} - \kappa) \xrightarrow{d} \mathcal{N} \left(0, \frac{\sigma_{\Delta_{\theta}}^2 - 2\kappa \sigma_{\Delta_{\theta} \Delta_{f_{\text{base}}}} + \kappa^2 \sigma_{\Delta_{f_{\text{base}}}}^2}{(e(f_{\text{base}}) - e(\mathcal{F}))^2} \right).$$

with $\sigma_{\Delta_{\theta}}^2$, $\sigma_{\Delta_{\theta} \Delta_{f_{\text{base}}}}$ and $\sigma_{\Delta_{f_{\text{base}}}}^2$ defined in the same way as in earlier sections.

0.5 How Different are Two Models?

The restrictiveness of parametric models does not tell us whether the two models rule out similar kinds of behaviors. For example, consider $\text{DA}(\eta)$ —which achieves a restrictiveness of 0.69—and $\text{CPT}(\gamma)$ —which achieves a restrictiveness of 0.59. Despite imposing different functional forms, these models may essentially capture the same risk behaviors, leading to their similar absolute levels of restrictiveness. Another possibility is that the two models embody rather different restrictions, so that mappings which are well approximated by $\text{DA}(\eta)$ are poorly approximated by $\text{CPT}(\gamma)$, and vice versa. We next provide a measure for determining whether two models are restrictive “in the same way.”

Consider two parametric models \mathcal{G}_1 and \mathcal{G}_2 , where we assume that $f_{\text{base}} \in \mathcal{F}_{\mathcal{G}_1}, \mathcal{F}_{\mathcal{G}_2}$.

For an arbitrary mapping f , define

$$\delta_f^1 := \frac{d(\mathcal{G}_1, f)}{d(f_{\text{base}}, f)} \quad \delta_f^2 := \frac{d(\mathcal{G}_2, f)}{d(f_{\text{base}}, f)}$$

to be the normalized discrepancy between the model and f .

Definition O.2. The δ -correlation between models \mathcal{G}_1 and \mathcal{G}_2 is the correlation coefficient for the pair (δ_f^1, δ_f^2) where f follows a uniform distribution on the admissible set \mathcal{F} .

Two models with a high δ -correlation do relatively well on the same mappings, while the δ -correlation is negative if the mappings that one model approximates well are relatively harder for the other to approximate.

The size of δ -correlation between two models does not directly imply anything about their absolute levels of restrictiveness.⁴⁵ The size of δ -correlation also does not tell us which model is more restrictive. If two models perform better on the same mappings, but one model fits all mappings better than the other, the δ -correlation measure will not reveal which of the two models is more flexible. The measure of δ -correlation, however, can be usefully paired with the restrictiveness of the two models to provide further insight into their comparison, as we now demonstrate.

O.5.1 Certainty Equivalents

Table 8 reports the δ -correlation between the models CPT- γ , CPT- (δ, γ) , CPT- (α, δ, γ) , DA(η), and DA(α, η).

	CPT(γ)	CPT(δ, γ)	CPT(α, δ, γ)	DA(η)	DA(α, η)
CPT(γ)	1	0.38	0.26	-0.76	0.40
CPT(δ, γ)	-	1	0.95	0.37	0.43
CPT- (α, δ, γ)	-	-	1	0.48	0.40
DA(η)	-	-	-	1	0.47
DA(α, η)	-	-	-	-	1

Table 8: δ -correlation between various pairs of models

⁴⁵For example, the models $\mathcal{G}_1 \equiv \mathcal{F}$ and $\mathcal{G}_2 \equiv \mathcal{F}$ have a δ -correlation of 1, as do the models $\mathcal{G}'_1 \equiv \{f_{\text{base}}\}$ and $\mathcal{G}'_2 \equiv \{f_{\text{base}}\}$. But the first pair of models is maximally unrestrictive while the second is maximally restrictive.

$\text{CPT}(\delta, \gamma)$ and $\text{CPT}(\alpha, \delta, \gamma)$ are highly correlated. Since the two models have similar absolute levels of restrictiveness ($r = 0.28$ for $\text{CPT}(\alpha, \delta, \gamma)$ and $r = 0.37$ for $\text{CPT}(\delta, \gamma)$), this suggests that the two models rule out very similar behaviors.

The two models $\text{DA}(\eta)$ and $\text{CPT}(\gamma)$ also have similar absolute levels of restrictiveness ($r = 0.69$ for $\text{DA}(\eta)$ and $r = 0.59$ for $\text{CPT}(\gamma)$). But their δ -correlation turns out to be quite negative, suggesting that the two models perform relatively well on different mappings. The models are thus different in empirical content and not simply in the statement of their functional forms. Interestingly, the gap in restrictiveness between $\text{CPT}(\gamma)$ and $\text{DA}(\alpha, \eta)$ is not substantially larger, but the δ -correlation between these models rises to 0.40, suggesting that introduction of the α parameter in addition to the η parameter in DA re-directs the model’s predictions in the direction of $\text{CPT}(\gamma)$.

The remaining δ -correlations are all positive but not large, suggesting that there are substantial differences between the models. The imperfect correlation is not surprising, since these model pairs are differentiated in both restrictiveness and completeness.

O.5.2 Initial Play

Table 9 compares the δ -correlation between models PCHM, Logit Level-1, and Logit PCHM.

	PCHM	Logit Level-1	Logit PCHM
PCHM	1	0.67	0.77
Logit Level-1	-	1	0.94
Logit PCHM	-	-	1

Table 9: Correlation between errors of the two models

The δ -correlation between Logit PCHM and Logit Level-1 is close to 1, so the distributions that these models fit relatively better and relatively worse are very similar. Since the absolute levels of restrictiveness for the two models are not statistically different, the near-perfect correlation in errors suggests that the two models have approximately the same empirical content. In contrast, PCHM—which is less complete and more restrictive than both Logit Level-1 and Logit PCHM—has a lower

δ -correlation with each of these models, although PCHM’s δ -correlation with Logit PCHM is slightly higher. This reflects the fact that the predictions of PCHM are more similar to the predictions of Logit PCHM than to the predictions of Logit Level-1.

O.6 Analytical SE with Clustering

We discuss here an alternative method for calculating clustered standard errors for completeness. The analytic standard errors we find here are very similar to the bootstrapped standard errors we reported in Appendix C.

We consider each subject’s reported certainty equivalents for the 25 lotteries as a 25-dimensional vector. We assume that this 25-dimensional vector is i.i.d. across subjects, but leave the dependence within this subject-specific vector unrestricted. Specifically, define the feature space \mathcal{X} to be a singleton consisting of the 25×3 matrix whose rows are the different lottery tuples $(\bar{z}, \underline{z}, p)$ in the Bruhin et al. (2010) data. The outcome space is $\mathcal{Y} = \mathbb{R}^{25}$, where a typical element is a vector of 25 certainty equivalents for the 25 lotteries. The expected certainty equivalent vector over subjects is represented by a mapping $f : \mathcal{X} \rightarrow \mathbb{R}^{25}$, which is simply a vector in \mathbb{R}^{25} .

Finally, let the loss function l be

$$l(f, Y_i, X) := \frac{1}{25} \|Y_i - f_\theta(X)\|^2 = \frac{1}{25} \sum_{h=1}^{25} (Y_{i,h} - f_h)^2.$$

This loss function groups together the squared losses of each individual subject across the 25 lotteries. Under this setup, the analytical formula for standard errors provided in Section 4.2 and Appendix B.2 can be directly applied, with sample size $N = 179$. Table O.6 reports the standard errors for completeness computed in this way.

	# Param	Completeness
CPT Specifications		
α, δ, γ	3	0.95 (0.09)
δ, γ	2	0.95 (0.08)
α, γ	2	0.95 (0.09)
α, δ	2	0.27 (0.09)
α	1	0.25 (0.05)
δ	1	0.26 (0.06)
γ	1	0.71 (0.06)
DA Specifications		
α, η	2	0.27 (0.06)
η	1	0.27 (0.05)