Evaluating and Extending Theories of Choice Under Risk*

Drew Fudenberg and Indira Puri

Department of Economics, MIT

First posted: February 2, 2021
This version: January 17, 2022

Abstract

We evaluate how well economic models of choice under risk predict the certainty equivalents of lotteries with known probability distributions. We use a new dataset that includes lotteries with varying numbers of outcomes, so we can test if lotteries with a smaller number of outcomes are preferred, as *simplicity theory* (Puri (2020)) predicts. A heterogeneous agent model that combines simplicity theory with cumulative prospect theory has the highest outsample predictive accuracy, and comes close to matching machine learning performance, which suggests that the model captures most of the regularities in the data. The model uncovers three behavioral types in the population, two complexity averse and one complexity loving, each with different levels of probability weighting. Demographic characteristics do not predict behavioral type, but financial literacy does. The model predicts well both on experimental data from prior papers and on real-world takeup of nonstandard savings accounts.

*We thank Monica Agrawal, Yoram Halevy, Gustav Karreskog, Annie Liang, and Charles Sprenger for helpful comments, and NSF Grant 1951056, the NSF Graduate Fellowship, and the Paul and Daisy Soros Fellowship for New Americans for financial support.

1 Introduction

Risk preferences play a significant role in many economic decisions, so it is important to model them accurately. Towards that end, we evaluate both classic and modern alternatives to expected utility theory. We also compare their performance to that of machine learning algorithms to see if there are predictable regularities that the economic models do not fully capture.

The most commonly used alternatives to expected utility theory are prospect theory (PT, Kahneman and Tversky (1979)), which allows for non-linear probability weighting and cumulative prospect theory (CPT, Tversky and Kahneman (1992)), which adds rank dependence to PT. More recently, simplicity theory (Puri (2020)) posits that people assign a utility premium to lotteries with fewer possible outcomes. This is consistent with, but stronger than, a preference for deterministic payments over stochastic ones.

We evaluate these models and combinations of them, which we term PT-Simplicity and CPT-Simplicity. We then compare the outsample predictive performance of these theories and that of machine learning algorithms on a dataset where each participant faces lotteries with a range of two to six outcomes.[1] We evaluate the standard functional forms for utility functions and probability weighting, and a functional form for simplicity theory that we introduce here. For each model, we allow for heterogeneous agents. For example, if the model is expected utility, we allow different groups of people to have different risk aversion parameters; if the model is PT, we also allow people to have different levels of probability weighting; and so on.

We find that a heterogeneous-agent model combining simplicity theory with CPT (the CPT-Simplicity model) does best. This model comes close to matching machine-learning performance, which suggests that it captures most of the empirical regularities in the data; the other models all perform substantially less well. PT performs worse than all models except expected utility; it is outperformed by both simplicity theory alone and CPT alone.

The heterogeneous CPT-Simplicity model uncovers three behavioral types in the population. Group 1, which comprises about 55% of our data, is complexity averse and distorts probabilities mildly; we call this group 'less behavioral, complexity

[1] By outsample, we mean the standard definition of the term: we train our models on one dataset, and test them on another. The performance on the unseen data is our metric.

averse'. Group 2 (the 'more behavioral, complexity averse' group), comprising about 15% of our data, is also complexity averse but distorts probabilities strongly. Group 3 (the 'complexity loving group') consists of about 30% of people. It is complexity loving and has nonlinear probability weights. We show that this third group is risk averse, illustrating the difference between liking complexity and liking risk.

We also test for whether participants prefer simplicity over and above any preference for certainty. We do so by calculating the residuals from the three-group CPT model, and regressing these on the number of outcomes. We find that the coefficient on number of outcomes is positive for all of the CPT groups, and statistically significant for one of them, which indicates a preference for simplicity over and above certainty.

We then test whether it is possible to predict a person's behavioral type from their observable characteristics. Income, sex, education, employment status, age, and race do not predict group membership in a statistically significant way. Financial literacy is predictive: a lower financial literacy score increases the probability of belonging to the minority complexity averse group. However, all three groups have substantial shares of both financially literate and financially illiterate individuals.

As tests of external validity, we apply the parameters and groups found by the heterogeneous CPT-Simplicity model to three other datasets. First, we show that the model accurately predicts the amount of Allais-type behavior in prior studies. Second, we show that our estimated model approximately replicates the event-splitting findings of Bernheim and Sprenger (2020), which compares two-outcome lotteries with three-outcome lotteries formed by adding a mean-preserving spread to what had been the more likely outcome.

Finally, we apply the model to predict real-world takeup of prize-linked savings (PLS) accounts. We show that the model makes accurate numerical predictions and that it also predicts the characteristics of those taking up PLS accounts. In particular, Group 3 has an outsize attraction to PLS, both because they overweight low probabilities, and because they assign a utility premium to lotteries with more outcomes. Group 2 drastically overestimates the probability of winning a prize, so although they are complexity averse, they are also drawn to PLS some of the time. Similarly, the less behavioral Group 1 balances an aversion to more outcomes against a mild degree of probability weighting. In our simulations, about half of those taking up PLS accounts belong to the complexity loving group (Group 3), with the remaining half split

between the two complexity averse groups. This implies that while no demographic characteristic may be able to predict the takeup of PLS accounts, at least half of those taking up PLS accounts should be those who would play lotteries anyway. Both our numerical predictions as well as these characteristic-based predictions are borne out by the data.

Our estimation technique for heterogeneous economic models improves on earlier work in several ways, which allows it to more consistently find parameters used to generate synthetic data. For example, we use a validation set to choose the number of groups and other hyperparameters. We also use a two-step process for estimating the parameters of each model, which can help in finding a global rather than local maximum.

We focus on pure economic models, rather than on hybrids of machine learning and economic theory, because pure economic models are easier to interpret and also easier to port to other domains to evaluate comparative statics and policy implications.

2 Literature Review

Our paper relates to work on models of choice under risk, mixture models, machine learning, and ways to compare the performance of theories.

Choice Under Risk The literature that directly tests (and typically rejects) the independence axiom of expected utility theory includes studies of the Allais paradox (including Conlisk (1989), Fan (2002), Huck and Müller (2012), Mongin (2019)) and also tests of the independence axiom with various combinations of probabilities and support sizes (Harless and Camerer (1994)). Fehr-Duda and Epper (2012) points out that probability weighting has been repeatedly observed in the lab, though evidence outside the lab is mixed. Most of the work that estimates or tests CPT focuses on two-outcome lotteries only (Tversky and Kahneman (1992), Tversky and Fox (1995), Wu and Gonzalez (1996), Abdellaoui (2000), Bleichrodt and Pinto (2000), Booij, van Praag and van de Kuilen (2010), Tanaka, Camerer and Nguyen (2010)). Our dataset extends beyond two-outcome lotteries, to include lotteries with many possible outcomes. We also test PT against CPT, to see which has better outsample performance; we are not aware of other work that does this.

Although PT and CPT incorporate probability weighting, and help explain many

experimental results, their validity is challenged in several recent studies (Etchart-Vincent (2009), Andreoni and Sprenger (2011), Bernheim and Sprenger (2020)). Even for papers that find evidence of PT or CPT behavior, parametric values vary widely from study to study (Neilson and Stowe, 2002).

Some part of the observed violations of PT and CPT seems to reflect a preference for certainty; as summarized by Harless and Camerer (1994), "in the triangular interior [e.g for lotteries with three outcomes], however, EU manages a miraculous recovery." Dillenberger (2010) captures a preference for certainty with the axiom of negative certainty independence, and Cerreia-Vioglio, Dillenberger and Ortoleva (2015) uses this axiom to provide a representation theorem for "cautious expected utility."

Our paper is also related to past work on a preference for fewer outcomes. Puri (2020) introduces and axiomatizes a theory of simplicity, which posits that agents incur a weakly increasing disutility from increases in a lottery's support, providing a novel theoretical foundation with which to understand and tie together disparate experiments across several literatures. The empirical evidence so far suggests that a preference for simplicity may explain regularities in behavior that PT cannot: Bernheim and Sprenger (2020) finds evidence for simplicity prone behavior using an event splitting task; Goodman and Puri (2020) finds that the behavior of retail traders in binary options markets cannot be rationalized by PT, but is predicted by simplicity theory; Mo□att, Sitzia and Zizzo (2015) and Sonsino, Benzion and Mador (2002) find that, holding fixed the expected value, people prefer lotteries with fewer outcomes even when those lotteries have higher variance. Because simplicity theory appears to explain regularities that other theories cannot, we utilize our framework to explore whether combining PT with simplicity theory produces a model as powerful as machine-learning algorithms. If the latter does not hold, we aim to explore which, if any, regularities in choice under risk both models miss.

Mixture Models Andersen, Harrison and Rutström (2006), Harrison and Rutström (2009), and Harrison, Humphrey and Verschoor (2010) use finite mixture models, but rather than a mixture of individuals, these papers use a mixture model on choices. Conte, Hey and Mo□att (2011) uses a mixture model that assumes one group of individuals behaves according to expected utility. Bruhin, Epper and Fehr-Duda (2010) estimates a mixture model for PT/CPT on binary lotteries, where these two theories are equivalent. It notes that "Unfortunately, no single best fitting model has

been identi ed so far... What [applied economics needs] is a parsimonious representation of risk preferences that is empirically well grounded and robust. Fudenberg et al. (2020) also estimates CPT with three types of agents on binary lotteries, and nds that it does a good job of predicting the average certainty equivalents of each lottery.

In addition to using models besides CPT, our mixture model improves on the optimization and assessment procedure. We use a validation set to pick the number of groups, so that this choice is based on the performance on unseen data; and we test model performance on unseen test data, rather than the training set. We also introduce a two-step process to help nd a global maximum, several xes for numerical stability, and analyze non-parametric behavior for each group found in the mixture model. We also analyze the the relationship between nancial literacy, demographic data, and group membership.

Machine Learning and Risk Preferences Erev et al. (2010), Erev et al. (2017), Plonsky et al. (2016), and Plonsky et al. (2017) run prediction contests in which participants were invited to submit machine learning algorithms to predict people's choices over lotteries. Plonsky et al. (2016) and Plonsky et al. (2017) nd that a combination (representative agent) ML-theory algorithm performs better than other contest entries. Ke et al. (2020) similarly introduces a combination ML-theory algorithm. Our focus is instead on whether any economic theory can predict choices well; we do not introduce or evaluate combination ML-theory algorithms. Also, the models in these papers are implicitly representative agent: the same algorithm is applied to everyone. In contrast, we allow for heterogeneous agent models. Finally, when these papers use non-binary lotteries, they are lotteries with either a binomial distribution, or a left- or right-skewed binomial distribution. Our multi-outcome lotteries are generated uniformly at random, which gives these lotteries more variety both in the outcomes and in the probabilities assigned to those outcomes.

Peysakhovich and Naecker (2017) uses ridge regression to predict choices on binary lotteries, and compare its performance to CPT. We use a larger range of machine learning models, test a larger variety of economic models, use a dataset with more variance in the number of outcomes, and apply our ndings to external datasets. We contribute to this literature by using unsupervised learning techniques, which we nd perform best. We also use validation sets, rather than train or test sets, to select

hyperparameters.

Comparing the Performance of Theories Fudenberg et al. (2020) proposes that the completeness of a theory should be measured by comparing its percentage improvement over a naive model to that of a table-lookup algorithm. We modify this to a machine learning completeness score that uses a machine learning benchmark in place of a table-lookup benchmark.[2]

3 Experimental Design

3.1 Participants

To gather our data we ran a survey on Amazon Mechanical Turk from 9/28/2020-10/1/2020. Following standard practice, participants were cautioned that they would be paid if and only if they passed attention and comprehension checks. (See Online Appendix V for the experimental instructions). By the time the survey nished running, 200 people had passed attention or comprehension checks and 157, failing such a check, saw a message informing them of that fact. These ltering numbers are in line with prior work using Mechanical Turk (Hauser and Schwarz (2016), Abbey and Meloy (2017), Chmielewski and Kucker (2020)). We further lter participants by dropping those who always provided the highest possible certainty equivalent for any lottery (the choice closest to the `next' button); this removes four people, who tended to take less time to answer the survey than others.[3] This gives us a nal sample size of 196. Participants were paid a $2.50 base rate, plus an incentive payment based on their response to a randomly chosen question. The average participant payment was $6.82, and the average time taken to complete the survey 19.14 minutes.

In addition to having participants provide certainly equivalents for lotteries as described in Sections 3.2 and 3.3, we ask participants about their income, age, sex, education, employment, and race. We also ask nancial literacy questions from Lusardi and Mitchelli (2007). There are three nancial literacy questions; the rst two are

[2] This is in the spirit of Peysakhovich and Naecker (2017) and Bodoh-Creed, Boehnke and Hickman (2018), which compare the performance of machine learning algorithms with that of models of risk preference.

[3] In Section III, we show that the results without dropping these four people, and with alternate forms of data cleaning, are similar to those presented in the body of the paper.

simple percentage calculations, and the third is a compound interest calculation (the exact questions are in Appendix A.2). We impose a time limit of 45 seconds to answer the first question, one minute to answer the second question, and five minutes to answer the third question. Following Lusardi and Mitchelli (2007), the participant receives a score of 2 if they answer the first or second question correctly and the third question correctly; they receive a score of 1 if they answer the first or second question, but not the third question, correctly; and they receive a score of 0 if they answer no questions correctly.

The summary statistics for each demographic and financial literacy question are shown in Table 6 in Appendix A.3. In our sample, 69% of participants receive the highest financial literacy score, 57% have at least a four-year college degree, and 72% are under the age of 40.

3.2 Random Lottery Generation

There are 50 lotteries, 10 for each of two, three, four, five, and six outcomes. We do not want there to be structural differences between lotteries with different numbers of outcomes, so we generate random lotteries as follows. Probabilities range from 0 to 1 and occur in increments of 0.01.[4] Payoffs range from \$1 to \$10 and occur in increments of 0.25. The range of the lottery can be any of 2, 2.5, 3, 3.5, or 4 dollars. By fixing the range before choosing the number of outcomes, we ensure that the difference between the least and greatest payoffs do not differ by number of outcomes. The full random lottery generation algorithm may be found in Appendix A.1, which also shows that the mean and variance of payoffs are not significantly correlated with the number of outcomes.

3.3 Obtaining Certainty Equivalents

We use the multiple choice list method, as in Bruhin, Epper and Fehr-Duda (2010), Bernheim and Sprenger (2020), and Chapman et al. (2017), among others.[5] To ensure

[4] This means that most probabilities are not integer multiples of .1. Huck and Weizsacker (1999)'s finding that the number of digits in probabilities or payoffs had no effect on behavior gives us some reason to hope that the same is true in our case.

[5] This is a standard way to determine risk preferences, but there is some evidence that other procedures yield different behavior, see e.g. Andreoni and Sprenger (2011), Freeman, Halevy and Kneeland (2019).

Figure 1: Sample Question

that participants do not have to spend excessive time clicking, we impose single switching, as in Andersen et al. (2006) and Tanaka, Camerer and Nguyen (2010).

Following Bruhin, Epper and Fehr-Duda (2010), for each lottery, each question has 10 evenly-spaced choices. The choices range from 50 cents below the lowest outcome in the lottery to the highest outcome in the lottery.[6] This allows participants to violate dominance if they choose, by picking a certainty equivalent below the lowest outcome in the lottery. Our lottery generation design ensures that the range of lotteries does not di er by number of outcomes, so the precision of the certainty equivalent obtained will not vary with the number of outcomes. The 50 lotteries were split into two sets of 25 each, and participants were asked to provide certainty equivalents for one of these sets. Figure 1 shows a sample question.

[6]To allow for evenly spaced steps, in some cases the highest choice was a few cents above the highest outcome in the lottery. The frequency with which this occurred was identical across di erent support sizes.

4 Preference Models

We consider monetary lotteries p with outcomes $0 < x_1 < x_2 < ::: < x_{\#p}$, where $\#p$ ranges from 2 to 6. Our basic preference models are expected utility theory, PT, CPT, and simplicity theory. The standard formulation of each model is listed below.

- Expected Utility Theory:

$$u(p) = \sum u(x)p(x)$$

- PT:

$$u(p) = \sum u(x)\ \ (p(x))$$

- CPT:

$$u(p) = \sum u(x_i) \left(\sum_{k=1}^{i} p_k \right) - \left(\sum_{k=1}^{i-1} p_k \right) ;$$

  where outcomes are in decreasing order (e.g. $k=1$ is the largest outcome in the lottery).

- Simplicity Theory:

$$u(p) = \sum u(x)p(x) \quad C(\text{jsupport}(p)\text{j})$$

  We also consider hybrid models that combine simplicity theory with either PT or CPT.

- PT-Simplicity:

$$u(p) = \sum u(x)\ \ (p(x)) \quad C(\text{jsupport}(p)\text{j})$$

- CPT-Simplicity:

$$u(p) = \sum u(x_i) \left(\sum_{k=1}^{i} p_k \right) - \left(\sum_{k=1}^{i-1} p_k \right) \quad C(\text{jsupport}(p)\text{j})$$

These hybrid models are proposed (but not axiomatized) in Puri (2020). The CRRA speci cation $u(x) = x$ is typical in empirical implementations of PT (Fehr-Duda

and Epper, 2012). We use the probability weighting function $w(p) = \frac{p^{\gamma}}{(p^{\gamma}+(1-p)^{\gamma})^{\frac{1}{\gamma}}}$ from Kahneman and Tversky (1979) and subsequent work.[7]

Simplicity theory is relatively new, so there is no standard functional form to use. We specify a three-parameter family of sigmoid functions, with $C(1) = 0$ as the theory requires: $C(x) = \frac{\alpha}{1+e^{-\beta(x-\delta)}} - \frac{\alpha}{1+e^{-\beta(1-\delta)}}$. In this specification, α represents the height of the function, and δ the midpoint of the rise, before normalization. The parameter β represents the slope, with higher β corresponding to a steeper slope. Because we normalize C so that $C(1) = 0$, larger values of α and δ also increase the height of the function, though to a lesser extent than increases in α.

When combining the models, we assume that the simplicity cost enters additively with the PT or CPT evaluation. There is no particular a priori reason that this should be the case, but as we will see this specification is very successful at predicting the certainty equivalents.

5 Econometric and Machine Learning Models

Instead of estimating the preferences of a representative agent, we use mixture models, as in Bruhin, Epper and Fehr-Duda (2010). A mixture model allows individuals to be classified into one of k groups, with each group having its own preference parameters. For example, an expected utility mixture model with k groups would allow for k different CRRA estimates, and each person would be mapped to one of these different estimates. This allows some groups of people to exhibit more probability weighting and other groups to exhibit less. Similarly, we can capture different risk preferences and responses to complexity. Our use of outsample evaluation for these models helps reduce the risk of overfitting; models with more parameters can have higher cross-validated prediction errors.

5.1 Evaluation

Each model maps parameter vectors to predicted certainty equivalents for each group and lottery. We evaluate models based on their mean-squared error (MSE), which is the average of the mean-squared error over all observations. We use cross

[7] This functional form is fairly standard, but some papers (e.g. Bruhin, Epper and Fehr-Duda (2010)) use a more flexible specification with an additional parameter.

validation to split our data into train, validation, and test sections.[8] The validation sets are used to set hyperparameters such as the number of groups, while the test sets are used to evaluate model performance on unseen data. Because we evaluate on outsample error, our results should generalize more broadly than if we evaluated on insample error, and models with more parameters do not necessarily have an advantage.

Unless otherwise stated, for all models below, we use the following cross validation procedure. Our data consists of 50 lotteries total, which we split into two sets of 25, A and B; each person provides certainty equivalents either for set A or set B. Within set A, we randomly divide lotteries evenly into s subsets, numbered $A_1; A_2; ...; A_s$. Within set B, we randomly divide lotteries evenly into s subsets numbered groups $B_1; B_2; ...; B_s$. We use A_1 and B_1 as test data, and A_2 and B_2 as validation data. The models are trained on all other groups. The model hyperparameters are chosen to be those under which the validation MSE is lowest given the parameters estimated on the train data. Once the model parameters and hyperparameters are determined, the model is evaluated on the test data. We use $s = 10$, so that our data is split into 80% train, 10% validation, and 10% test.

The above procedure constitutes one cross-validation split. To obtain more accurate parameter estimates, we use bootstrap (Efron and Tibshirani, 1993), repeating the above cross-validation procedure many times. (We specify the number of bootstrap iterations as part of the description of each machine learning or econometric model.)

5.1.1 Econometric Specification

We wish to estimate parameters for model m (for the list of models we estimate, please see Section 4). Denote by $\hat{ce}_{m,l}(\theta)$ the predicted certainty equivalent for lottery l when using model m with parameters θ. The observed certainty equivalent for individual i on lottery l is $ce_{i,l} = \hat{ce}_{m,l}(\theta) + \varepsilon_{i,l}$, where $\varepsilon_{i,l} \sim N(0, \sigma_{i,l})$, $\sigma_{i,l} = \sigma_i |range(l)|$, and we allow each individual to have their own variance σ_i in lottery selection. We multiply the variance by the range of the lottery because the precision of the certainty equivalent we obtain depends on the spread between the highest and lowest outcome (Section 3.3).

[8] Validation sets are used in machine learning to tune hyperparameters. A hyperparameter is a value that contributes to the description of the model, such as a regularization weight in Lasso.

Let $c = 1; 2; :::; C$ denote the different groups. Let $_c$ be the parameters associated with group c. The probability density of individual i's choices under parameters$_c$ is:

$$f(ce_i; _c; f _{i;l} g_{l2L}) = \prod_{l=1}^{L} \frac{1}{_{i;l}} \quad \frac{ce_{i;l} \quad \hat{ce}_{i;l}}{_{i;l}} \quad ;$$

where denotes the density of the standard normal distribution.

Let $_c$ denote the probability of group membership of type$c$. The log-likelihood of the nite mixture model is:

$$\sum_{i=1}^{N} \ln \sum_{c=1}^{C} {}_c f(ce_i; _c; _{i;l});$$

where the rst sum is over individuals and the second sum is over groups. The parameters to be estimated are$(_1; :: _C; _1; :::; _C; _1; :::; _N)$.

5.1.2 Estimation

We estimate the model using an expectation-maximization algorithm (Dempster, Laird and Rubin, 1977). This algorithm proceeds iteratively, rst calculating the log-likelihood of the model, and then estimating the parameters of the model using maximum likelihood. We repeat these two steps until the process converges.

With nite runs and nite data, care is required as our numerical algorithm may converge to a local maximum. To ensure that we reach a global maximum, we follow a two-step procedure: We rst iterate over 100 possible initializations, using 10-fold cross validation for each (e.g, bootstrap the cross-validation procedure 10 times for each initialization), and pick the best initialization based on validation set performance. We then bootstrap the cross validation procedure 100 times using this best initialization (see Appendix B.2 for further details).

We estimate parameters for$k = 1; 2;$ and 3 groups. Because the number of groups is a hyperparameter, we pick the best number of groups based on validation set performance. Our estimation procedure builds on that of Bruhin, Epper and Fehr-Duda (2010), with some changes: we evaluate based on outsample performance, and use validation performance to choose the number of groups for any given model, using the two-step procedure detailed in Section 5.1.2 rather than averaging over initializations. Also, because we have two levels of noise (the cross-validation procedure

and the initializations) as opposed to one (just the initializations), we add several fixes for numerical stability: We allow α to depend on g in addition to i, impose a lower bound on each π_k of $(\text{number of groups})/100$, and impose a lower bound on $\gamma_{i,g}$ of $1/100$ and an upper bound of 100. These fixes for numerical stability allow the mixture model to more consistently find parameters used to generate synthetic data. Finally, as demonstrated by simulation in Appendix B.1, the log likelihood of non-data-generating parameters increases as the number of observations per lottery decreases. With 200 people divided into 3 groups, at least one group may be small, and without constraining r, some runs result in implausibly low levels of r, e.g. 0.2 or 0.3. To address this issue, we impose lower and upper bounds on r, which are guided by the literature. We conduct robustness checks for alternative bounds in Appendix III.III.

5.2 Machine Learning Algorithms

We predicted certainty equivalents using three different machine learning algorithms: neural networks, k-means, and gradient boosting trees. Here we describe how we implemented these algorithms.

5.2.1 Neural Networks

Our output variable for the neural network is the observed certainty equivalent for each lottery and individual. To obtain stronger performance, we allow the neural network to use the following sets of input variables: {the set of outcomes, the set of probabilities, and an indicator for each individual}; {expected value, variance, number of outcomes, and an indicator for each individual}; {expected value, variance, and an indicator for each individual}; and each of these without the individual-specific indicators. Similarly, we allow for any of the following architectures: 2-layer, 2-hidden unit; 2-layer, 3-hidden unit; and 3-layer, 2 hidden units per layer, choosing the best variable set by validation set performance. Having chosen the best input variable set and architecture using validation performance, we report test performance in the results section (Section 6). For each input variable set - architecture pair, we evaluate using 10-fold cross validation and 30 different initializations.

5.2.2 K-Means

K-Means is an unsupervised algorithm that clusters similar groups together. To apply it, we use a cross-validation procedure similar to that described in Section 5.1. Lotteries are split into train, validation, and test sets. We group individuals into k groups using the train data. On the validation data, we predict individual i's certainty equivalent for lottery g to be the average of the certainty equivalents for lottery g picked by everyone else in individual i's group.[9] Rather than using the full 50 lotteries all at once, here we are forced to split them into two sets of 25, because the clustering happens on a lottery basis, and each individual answers only 25 lotteries. For this reason, to ensure a reasonable sized validation and test set, in this algorithm we split our data into 60% train, 10% validation, and 30% test.[10] We attempt up to ten groups, and pick the best group number and initialization based on validation set performance. For each group, we evaluate using 20-fold cross validation across 250 different initializations.

5.2.3 Gradient Boosting Trees

A gradient boosting tree uses multiple splits of the data to predict certainty equivalents. As in Section 5.2.2, we use the algorithm separately on the first 25 and second 25 lotteries, since the standard algorithm is sensitive to missing data. Consequently, we split our data into 60% train, 10% validation, and 30% test in each bootstrap iteration. The input variable set is {expected value, variance, number of outcomes, and an indicator for each individual}, and the hyperparameter we focus on is the maximum depth of the tree. We allow for a maximum depth of 1,2,3,4,5, and 10, choosing the best depth and initialization by performance on the validation set. For each maximum depth, we evaluate using 20-fold cross validation across 250 different initializations.

[9] To ensure the validation and test predictions are well defined, we impose the constraint that the minimum group size is two. For this, we use the standard Python implementation, which is based on Bradley, Bennett and Demiriz (2000).

[10] Using 80% train, 10% validation, and 10% test results in extremely small validation and test sets here, and performs worse than using 60% train.

6 Analysis

6.1 Relative Scores

Table 1 compares test performance of the economic theories to that of the best-performing machine learning algorithm, which turned out to be k-means; every model performs best with three groups.[11] We compute a score for each model, similar to Fudenberg et al. (2020). That paper suggests calculating a completeness score by using the performance of a naive algorithm and a benchmark hold-one-out evaluation method. Here, we use expected value as our naive algorithm, but as a benchmark we use machine learning, so we report the machine learning completeness score

$$\text{Score (model)} = \frac{MSE_{naive} - MSE_{model}}{MSE_{naive} - MSE_{ML}}$$

We emphasize that we measure completeness with respect to this dataset and our machine learning algorithms. A larger dataset or one with more features could improve the machine learning benchmark, as could advances in machine learning algorithms. Conversely, we might expect larger prediction errors for both machine learning and for the CPT-Simplicity model on a dataset with more variation in the magnitudes of the prizes.

The combination model CPT-Simplicity achieves a 96% ML completeness score when used with three groups, which turns out to be optimal.[12] The next best models, in order, are: CPT, PT-Simplicity and Simplicity (tied), PT, and EU. Simplicity on its own outperforms PT, and it performs about as well as the PT - simplicity model. CPT, while performing better than PT, still misses regularities in behavior.[13] Combining simplicity with CPT works well.[14]

[11] See (Table 11 in Online Appendix I). The best performance of each machine learning algorithm, as measured by validation set MSE, is reported in Table 12 in Online Appendix I). The fact that the best performing machine learning algorithm is k-means may re ect the size of our dataset with millions of observations, we might expect a neural network to perform better.

[12] On the validation set, three groups gives a 20% improvement in MSE relative to using one group.

[13] Because of our more complex set of lotteries, CPT ts less well here than in the binary lotteries studied by Bruhin, Epper and Fehr-Duda (2010) and Fudenberg et al. (2020), both in an absolute sense and compared to ML performance.

[14] We emphasize that this does not mean the CPT is the right way to model non-linear probability weighting, but the fact that CPT-Simplicity outperforms PT-Simplicity suggests that rank dependence is a useful component of predictive models. Note also that the fact that PT-Simplicity does much worse than CPT-Simplicity shows that the PT-Simplicity model is at least somewhat restrictive, which suggests that CPT-Simplicity may be as well.

Of the three CPT-Simplicity groups, both Group 1 and Group 2 are complexity averse, Group 2 more so than Group 1; Group 3 is complexity loving. This finding is consistent with Moffatt, Sitzia and Zizzo (2015), which finds that 23% of subjects are complexity loving; in our data, that number is about 30%, with the remaining 70% belonging to one of the two complexity averse groups. The standard errors on most parameters are relatively low. Because the standard errors on the simplicity parameters in our less behavioral complexity averse group are higher, we perform further tests (Section 7.2) to determine whether this group exhibits complexity averse behavior. These tests show that this group is indeed complexity averse (Section 6.2). The standard errors here may reflect that the sigmoid function may not be the most appropriate simplicity function for this group.

Figure 3 graphs the simplicity functions for each group. Although we call the complexity averse groups more or less behavioral based on their probability weighting, both groups display a preference for simplicity, in different ways. Both groups are averse to even two outcomes, but complexity aversion is stronger in the more behavioral group. The complexity loving group enjoys uncertainty; their enjoyment of uncertainty does not increase in the number of outcomes. Both complexity averse groups display an aversion to more outcomes, with complexity aversion stronger in the more behavioral group. The complexity loving group enjoys lotteries with two or more outcomes, but within a fixed number of outcomes are risk averse (Online Appendix I.III). The CRRA parameter for all groups is around 0.8 - 0.9, consistent with estimates in prior experimental work using prospect theory (Kahneman and Tversky (1979)), Bruhin, Epper and Fehr-Duda (2010)).

Table 1: Test Performance

Model	Test MSE	ML Completeness Score
Expected Value	82.87	0%
	(12.26)	
EU	80.18	15.05%
	(12.00)	
PT	72.95	55.51%
	(10.60)	
Simplicity	70.81	67.49%
	(10.75)	
PT-Simplicity	70.78	67.66%
	(10.45)	
CPT	69.18	76.61%
	(10.21)	
CPT-Simplicity	65.57	96.81%
	(10.39)	
Machine Learning Benchmark	65.00	100%
	(6.88)	

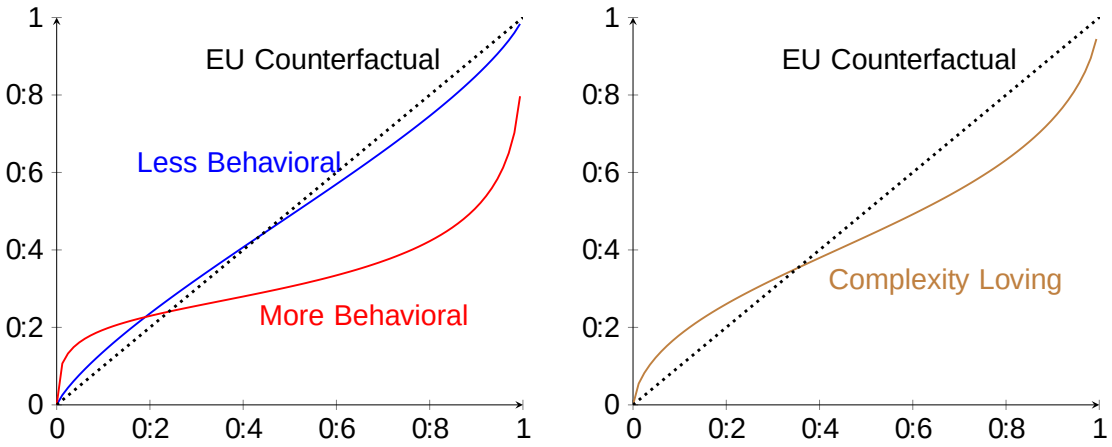
Standard deviation in parentheses.

Table 2: CPT-Simplicity Group Parameters

	Complexity Averse		Complexity Loving
	Less Behavioral	More Behavioral	
	0.794	0.803	0.884
	(0.052)	(0.056)	(0.096)
	0.828	0.441	0.642
	(0.030)	(0.020)	(0.022)
	1.974	0.716	0.735
	(2.822)	(1.572)	(0.323)
	2.874	2.927	4.659
	(3.018)	(2.169)	(1.410)
	0.344	1.216	-2.606
	(0.580)	(1.093)	(2.398)
Number	100.650	30.920	64.430
	(5.887)	(3.148)	(4.672)

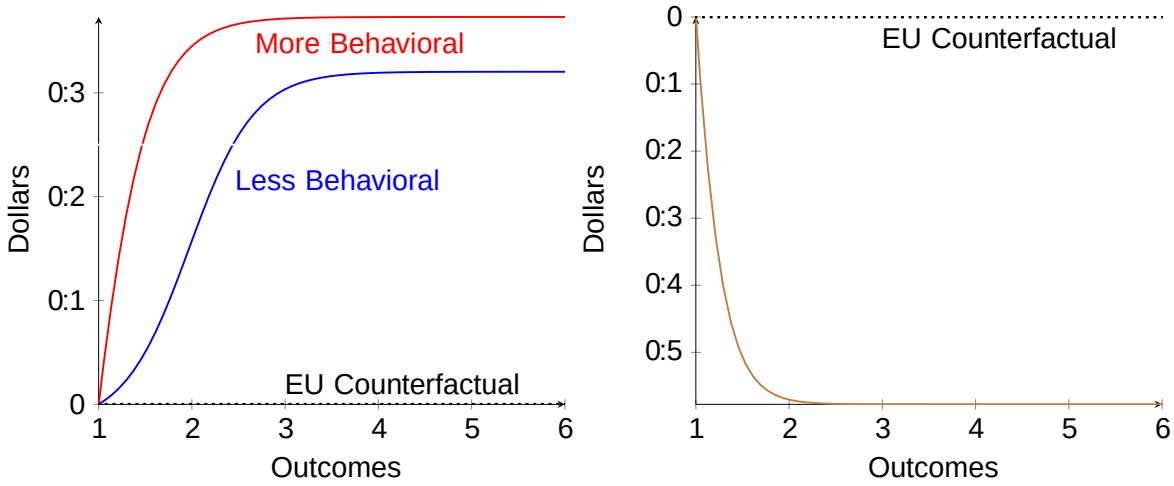
Standard deviation in parentheses.

Figure 2: Probability Weighting Functions By Group



(a) Complexity Averse (b) Complexity Loving

Figure 3: Simplicity Functions By Group



(a) Complexity Averse (b) Complexity Loving

6.2 Group Composition

The propensity scoreof individual i for group c is the average of the probability that individual i belongs to groupc, taken over all bootstrap iterations. Figure 5 in Appendix A.4 shows that the propensity scores are near zero or one for most individuals and groups, so there is little ambiguity about the group to which an individual belongs.[15]

We now statistically test whether certain characteristics predict group membership. In each bootstrap iteration of the mixture model, we calculated the probability that an individual belonged to a particular group. For each group, we regress the probability that an individual belongs to group c in bootstrap iteration b on the matrix of nancial literacy and demographic characteristics for that individual. We cluster standard errors at the individual level.[16] Table 10 in Appendix C shows that most demographic characteristics - age, employment status, gender, race cannot predict group membership at the 5% signi cance level. Few variables are predictive, and even those that are explain little of the variation in group composition. The only

[15] In addition to the completeness to non-parametric machine learning algorithms, the cleanness of classi cation provided another way of measuring the suitability of the parametric model. Halevy, Persitz and Zrill (2018) provides a di erent, in-sample method of nding parameters with the goal of minimizing deviations from the chosen functional form.

[16] Precisely, the speci cation is $p_{i;c;b} = X_i + _{i;c}$, where $p_{i;c;b}$ is the probability that individual i belongs to groupc in bootstrap iteration b and x_i are individual characteristics.

variable that has predictive content is the financial literacy score: Individuals who are less financially literate are statistically significantly more likely to be in the more behavioral, complexity averse group (Group 2). A decrease in the financial literacy score of one unit increases the probability of being in this group by 13%.

Several effects have marginal statistical significance: higher financial literacy increases the probability of belonging to the less behavioral, complexity averse group (Group 1); lower income increases the probability of belonging to the more behavioral, complexity averse group (Group 2); college education increases the probability of belonging to the more behavioral, complexity averse group; and being white increases the probability of belonging to the complexity loving group (Group 3).[17]

Some of these findings are broadly consistent with the literature: low income and financial literacy have been associated with more deviations from expected utility (Xiao (2008), Lusardi and Mitchell (2014)). However, it is worth noting that none of age, sex, race, education, or employment can predict probability weighting or complexity aversion at a statistically significant level in our sample.

7 Data Visualization

In this section we examine risk-taking behavior, probability weighting, and complexity aversion without making any parametric assumptions. To examine probability weighting, we follow Bruhin, Epper and Fehr-Duda (2010) and plot the median relative risk premium $(EV - CE)/EV$.[18] Probability weighting predicts that the relative risk premium is weakly increasing in the probability of the better outcome. We also plot the risk premium against the variance of the lotteries to examine attitudes towards risk.

[17] A one unit increase in financial literacy increases the probability of belonging to the less behavioral complexity averse group by 9% (p-value 0.12). A decrease in one's reported income bucket by one unit (which corresponds to $25,000 on average) increases the probability of belonging to the more behavioral complexity averse group (Group 2) by 2%. Being college-educated increases the probability of being in the more behavioral complexity averse group (Group 2) by 9%, but this effect is not significant at the 5% level). Being white increases the probability of belonging to the complexity loving group (p-values 0.13).

[18] While we follow the literature in calling the object $EV - CE$ the risk premium against the probability of obtaining the better outcome in two-outcome lotteries (Fehr-Duda and Epper, 2012). Puri (2020) pointed out that in the presence of a simplicity preference, the risk premium is better thought of as a risk-complexity premium. Indeed, Online Appendix I.III shows by example that the so-called `risk premium' captures both risk aversion and complexity aversion.

To examine complexity aversion, we use the CPT residuals, which are the differences between the actual certainty equivalents and those predicted by CPT with our estimated parameters, divided by the range of the lottery. Note that this is the standard normal residual from our econometric model (up to a constant). If CPT is the true model this residual should be unrelated to the number of outcomes, whereas complexity aversion predicts the residual would be increasing in the number of outcomes. In all regressions that follow, we include an intercept term, so that any relationship between number of outcomes and CPT residuals is over and above both what CPT, and what CPT plus a certainty premium, would predict.

7.1 Overall

Figure 6 in Online Appendix I.II shows that individuals respond strongly to expected value. They respond mildly to risk, with the risk premium broadly increasing with variance. There is some evidence of probability weighting, but it is not strong: the median risk premium is negative for probabilities below 0.3, and mostly positive afterwards. However, contrary to the prediction of probability weighting models, the relative risk premium does not appear to be increasing in the probability of the better outcome. There is some evidence for complexity aversion: a regression of the CPT residual on the number of outcomes, clustering standard errors by individuals, yields a marginally significant coefficient of .0037 (p-value 0.108).

A representative agent here would be moderately complexity averse and display little probability weighting. We had instead fit a heterogeneous agent model. The next few subsections consider the unconditional data by group, checking whether the data without parametric assumptions maps back to the parametric behavior we found. Breaking the data by group allows more nuanced patterns to emerge.

7.2 By Group

We show that the non-parametric behavior described for each group is qualitatively similar to the parametric behavior predicted for that group. We assign an individual to a group if their propensity score for that group is at least 0.6. This procedure successfully assigns 178 of the 196 individuals (about 90%) to groups.

In additional to the graphical analysis described earlier, we also examine complexity aversion analytically, by regressing the CPT residual against the number of

outcomes, clustering standard errors at the individual level. To construct a CPT residual for each group, we use the and estimates for that group as found using a representative agent CPT model for that group.

Figure 7 shows the probability weighting, variance, and expected value plots for Group 1 (all gures referred to in this section are located Online Appendix I.II). The group shows little evidence of probability weighting: the median relative risk premium is roughly increasing in the probability of the higher outcome for two-outcome lotteries, but not in a way that is obvious or large (Figure 7c). This group is only risk loving when the probability of the better outcome is below 0.2. The group shows evidence of mild risk aversion, with median risk premium slightly increasing in the variance of the lottery (Figure 7b).

This group shows evidence of complexity aversion: regressing the CPT residual on the number of outcomes, clustering standard errors by individual, yields a coe cient of 0.0144, with ap-value that rounds to 0:00.

Group 2 likewise displays evidence of complexity aversion, but also shows strong evidence for probability weighting. For complexity aversion, regressing the CPT residual on number of outcomes, clustering standard errors at the individual level, yields a coe cient of 0.0115, with p-value 0.157. Figure 8 shows there is a marked increase in the median relative risk premium as the probability of the better outcome increases for two-outcome lotteries (Figure 8c), which is non-parametric evidence of probability weighting (see Fehr-Duda and Epper (2012) and Bruhin, Epper and Fehr-Duda (2010)). This group responds more noisily to the expected value than the `less behavioral, complexity averse' group (compare Figures 7a and 8a). They also respond slightly more strongly to variance than the `less behavioral' group (compare Figures 8b and 7b).

Group 3 displays evidence for probability weighting (Figure 9c), as their relative risk premium is increasing in the probability of the better outcome. Consistent with the parametric results, which show a complexity preference that remains at away from certainty, regressing the CPT residual on number of outcomes, clustering standard errors by individual level, indicates that the amount by which this group values complexity does not increase in number of outcomes (coe cient -0.0011, $p = 0:87$). Like the other two groups, this group responds strongly to the expected value, with their noise on this measure falling between the two complexity averse groups (Figure

9a).[19]

7.3 Testing for a Simplicity Preference

Section 6.2 looked at the CPT residual, with the groups as found by the CPT-Simplicity model. This helped us understand behavior for each of those groups. To test for a simplicity preference, we now use the groups found by the CPT model[20], and construct CPT residuals for each of these groups. The null hypothesis behind this test is that the true model within each group is CPT, and that the CPT model correctly assigns people to groups.

Table 3 shows the coefficient and p-value after regressing the CPT residual on number of outcomes, clustering errors at the individual level. These regressions have a positive slope in all 3 groups; this slope is statistically significant for one group. The regressions also indicate complexity aversion over and above what CPT with a certainty preference can capture; in CPT with a certainty preference, we would expect the intercept to be positive and the slope flat, which is not what we find.[21]

An alternative explanation for the complexity aversion we find is that the probability weighting parameter γ is not fixed but varies with the cardinality of the lottery's support. To test this alternative theory, we hold γ fixed at the value found for each group in Table 2 (e.g, $\gamma = 0.8$), and separately estimate a CPT model for each number of outcomes for each group. Online Appendix II shows that the γ's do not vary monotonically by number of outcomes. Further, the test performance of varying γ's by number of outcomes, is less than the test performance of CPT-Simplicity, indicating that this alternative theory does not capture the behavior found by the CPT-Simplicity model.

[19] Complexity loving is distinct from risk aversion. Risk aversion implies that, on lotteries with the same number of outcomes, the agent has a higher risk premium for lotteries with more variance. Complexity loving means that more outcomes are preferred to fewer. Both behaviors are true of this group. Online Appendix I.III shows that on lotteries with a fixed number of outcomes, the median risk premium increases in variance, indicating that this group is risk averse.

[20] The parameters found by the heterogeneous agent CPT model are reported in Appendix I.IV.

[21] If the data is generated by CPT, then when we regress CPT residuals on number of outcomes without an intercept, there should be mean-zero residuals regardless of number of outcomes. However, the slope is statistically significant (p-values < 0.025) and positive (0.0065 and 0.0084 respectively) for two groups, and statistically significant and negative (slope -0.02, p-value 0.00) for one group.

Table 3: Regress CPT Residual on Number of Outcomes; Groups Found by Three-Group CPT Model

Group	Coefficient on N. Outcomes	p-value of Coefficient
Group A	0.0025	0.466
Group B	0.0075	0.056
Group C	0.0067	0.347

8 External Applications

In this section we examine the predictions of our estimated CPT-Simplicity model on data we did not collect. To make predictions for each group we draw parameters for that group from a normal distribution whose mean is the group's point estimate for that parameter and whose standard deviation is the group's standard deviation for that parameter as reported in 2.[22] We draw these parameters 1000 times. Some of the applications involve payoffs beyond the range of our experiment. In these cases we linearly decrease the payoffs to fall within the range that we examined.

8.1 Laboratory Data

The Allais paradox For the Allais paradox, the behavior our model predicts is close to those found in incentivized experiments. Conlisk (1989) reports that 6% of people violate EU in the Allais paradox; Huck and Müller (2012) find 8% of people display Allais behavior; and Harrison (1994) finds 15% of people displaying Allais behavior. Using the the payoffs of Conlisk (1989) (which are similar to those in Huck and Müller (2012)), our model predicts that neither the complexity loving group nor the less behavioral, complexity averse group display Allais behavior (in our simulations, these groups display Allias behavior 1.5% and 0.5% of the time, respectively), while the more behavioral, complexity averse group displays Allais behavior 24.4% of the time. Taking a population weighted average, our simulations predict that 9.5% of the population displays Allais behavior when faced with real, small stakes, in line with proportion of Allais behavior found by studies using these stakes.[23]

[22] This is standard Monte Carlo (Brandimarte, 2014) on group behavior, where the uncertainty arises from uncertainty over the parameter estimates. Here, individual heterogeneity is already accounted for by splitting the population into three groups.

[23] There is a vast literature on Allais with large and hypothetical stakes. We compare our predictions to the real stakes results because they are incentivized. The real stakes used in experiments

Neilson and Stowe (2002) notes that there is a large range of probability weighting estimates for CPT models. In particular, Kahneman and Tversky (1979) finds a CPT parameter of $= 0.61$; Camerer and Ho (1994) finds $= 0.56$; and Wu and Gonzalez (1996) finds $= 0.74$. Our heterogeneous agent model can accommodate the range of previous parameters found.[24]

Event splitting Our model also predicts the event splitting behavior found by Bernheim and Sprenger (2020), where spitting the probability of an event leads to a drop of the certainty equivalent that is not consistent with CPT. Paralleling their "split-low" case, we simulate the utility for each group for the lotteries $\{\$2, 0.6; \$3, 0.4\}$ and $\{\$2 - \varepsilon, 0.3; \$2 + \varepsilon, 0.3; \$3, 0.4\}$, where $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$.[25] Bernheim and Sprenger (2020) finds that experimental participants provide a lower certainty equivalent from the three-outcome lotteries, over and above what CPT would predict; and further the certainty equivalent is higher for the two-outcome lottery than for the three-outcome lottery with the smallest value of ε.[26]

Simulating behavior for each group, we find that the two complexity averse groups incur a utility penalty over and above what CPT would predict for the three-outcome lotteries relative to the two-outcome lotteries (see Figure 15 in Online Appendix IV.I), roughly replicating the empirical pattern found in Bernheim and Sprenger (2020). Taking a population-weighted average yields that the simulated certainty equivalent decreases by \$0.034, in line with Bernheim and Sprenger (2020). Figure 16 shows that we have analogous findings in the "split-high" case where the two-outcome lottery $\{\$2, 0.4; \$3, 0.6\}$ is compared to $\{\$2, 0.4; \$3 - \varepsilon, 0.3; \$3 + \varepsilon, 0.3\}$.

8.2 Prize-Linked Savings

In a PLS vehicle, rather than receiving an interest rate, the individual is entered into a drawing for a prize. The return from the drawing is typically less than the interest that would be earned on a comparable normal savings vehicle. PLS have

are also close to those used in our experiment.

[24] The lowest α we found in the three-group CPT-Simplicity model is around the lowest representative agent α in the papers discussed in this section; similarly, the highest α in the three groups roughly corresponds to the highest α from the representative agent literature.

[25] The payoffs they used are 10 times as large.

[26] It finds a difference of \$0.47 and with our 10x smaller stakes we find the corresponding value to be \$0.036.

principle guarantees, so no losses are involved. The institution providing these PLS vehicles may pro t either by borrowing at a lower rate than they would otherwise be able to (if the government) or by using the capital invested in the PLS to obtain a normal savings return (if a private entity). In this section, we brie y review the history of these products, and show that the CPT-Simplicity model accurately predicts PLS takeup.

8.2.1 Predicted Behavior

Suppose that the alternative to putting a dollar in a prize-linked vehicle is putting a dollar in a regular savings account. (In the South African case discussed in Section 8.2.2, these savings accounts are easily accessible to anyone using a PLS account.) Since the interest rate on a regular savings account is typically higher than the return on the PLS, the CPT-Simplicity model predicts that Group 1, which is less behavioral and complexity averse, will tend to prefer a regular savings account over a PLS. On the other hand, the complexity loving Group 3 may prefer the multi-outcome PLS to the single-outcome regular savings account. And Group 2, which is more behavioral and complexity averse, balances two countervailing forces: on the one hand, an aversion to more outcomes, and, on the other, a tendency to weight probabilities strongly so that they severely overestimate the chance of winning a low probability prize.

Because our experiment used payo s from $1 - $10, we rst convert to dollars, then linearly normalize payo s downwards to be those that would occur with a $5 minimum investment. We calculate the attractiveness of a single lottery draw versus the comparable product. The South Africa case we discuss below had a monthly draw, so we compare how attractive it would be relative to holding a normal savings product for one month. To obtain population predictions, we multiply the prediction for each group by the point estimate for the proportion of that group in the population. The reported standard deviation is calculated using the same method. Note that this does not account for the noise in the point estimate of the proportion of the group in the population, so that the standard deviations we report are a lower bound for the model's predictions.

8.2.2 South Africa

The prize-linked savings accounts in South Africa were termed `Million a Month', and were run by First National Bank, one of the largest retail banks in that country. The program ran from January 2005 through March 2008 before shutting down due to legal challenges from the South Africa Lottery Board. Cole, Iverson and Tufano (2021), who had access to the First National Bank data, reports that by March 2008, the monthly prize amounts were: one prize of R1,000,000; four prizes of R100,000; 20 prizes of R10,000; and 200 prizes of R1,000. An investment of R100 gave a person one entry into the prize drawing. The same paper also reports that by 2008, the as-if APR on a prize-linked account was 1.81% and was stable; they suggest that this as-if APR should be considered the `equilibrium value' for the prize-linked account return. As of November 2004, for balances below R10,000, the standard savings account paid 4% annual interest; for balances between R10,000 and R25,000, it paid 4.25% APR; and for balances from R25,000 to R250,000, the APR ranged from 4.5% to 4.75%. The average exchange rate from USD to Rand during this period was 6.78 FRED (2020). Using these numbers, we use the calibration method described in Section 8.2.1 to obtain predictions from the CPT-Simplicity model.

The result of the calibration is shown in Table 4. There are two forces to consider: the low probability of a very high prize, and the fact that the prize-linked lottery has ve outcomes. Group 3 both overweights low probabilities and likes complexity, so they have 96% take up the prize-linked account in our simulation. Group 2 dislikes lotteries with more outcomes, but weights low probabilities extremely highly, so 98% of the group take up the product in our simulations. Finally, Group 1 has more linear probability weights and dislikes more outcomes, so most of them do not take up the product. Because they do still overweight low probabilities, though to a lesser extent than the other groups, a minority of this group (33%) take up the prize-linked account in our simulations. Overall, of those who take up the account, about half belong to the complexity loving group (Group 3). The remaining half are a mix of the complexity averse groups (Groups 1 and 2). Our model predicts that 64% of people overall will take up the prize-linked savings account.

Compare this to the percentage of First National Bank employees who take up the product, at 63% (Cole, Iverson and Tufano (2021)). Our model's prediction is within 1% of actual takeup. Similarly, during this period, 44.7% used a normal savings account, though as employees of the bank, they all had the ability to easily open

and use normal savings account. Our CPT-Simplicity model predicts that 64% of the population prefers PLS to normal savings, and the remaining 36% prefer normal savings to PLS: close to the 44.7% of employees who held a normal savings account during this period. Note that these takeup numbers are for the bank's employees rather than for the general population. In this setting, the prize-linked account was provided only by this bank and no other bank or state entity. Hence, there may have been barriers to advertising, knowledge of the account, or ability to open the account for the general population. For this reason, employee takeup is a useful proxy for the percentage that would have taken up an account with full access and knowledge, and it is the benchmark that we measure against.

Table 4: South Africa: Prize-Linked Savings by Group

	Complexity Averse		Complexity Loving	Total
	Less Behavioral	More Behavioral		
CPT-Simplicity	34%	98%	96%	64%
	(47)	(20)	(14)	(25)
CPT Only				84%
				(17)
Actual				63%

Standard deviation in parentheses.

Not only is actual takeup close to predicted takeup, the characteristics of those who take up these accounts also match our model. In particular, we found in Section 6.2 that none of: income, age, education, male, white, or employment predict group membership. So none of those demographic characteristics should predict PLS takeup either. Cole, Iverson and Tufano (2021), Table IV, reports precisely this: that none of income, white, age, or education predict PLS takeup.[27]

[27] First National Bank employees earn more, are less likely to be black, and more likely to have bank accounts relative to the average South African. In the table we refer to, the paper looks at demographic characteristics predicting takeup among the general South African population, by regressing log total PLS deposits at a branch on the demographic characteristics of the area the bank branch is located in. Regressing takeup on demographic characteristics for only bank employees (Table III of that paper) showed that the probability of having a PLS account does not vary much by income, race, or age, and that males are less likely than females to have any type of savings account at First National Bank.

Within those who take up the product, half belong to the complexity loving group. This group has an outsize attraction to lotteries, all else equal, because they prefer more outcomes. Our model therefore predicts that, while demographic characteristics do not predict takeup, those who take up prize-linked savings should largely be those who would otherwise still play lotteries. Cole, Iverson and Tufano (2021) presents evidence saying exactly this: that, while the demographic variables we report do not predict takeup, PLS substitutes for playing in the state lottery.

Finally, note that simplicity plays an important role in obtaining an accurate prediction: the three-group CPT model (without simplicity) incorrectly predicts almost universal takeup (84%), because without a simplicity preference, the prize lottery is attractive to any group that su ciently overweights small probabilities.

9 Conclusion

We used machine learning techniques as a benchmark against which to measure the performance of theories under risk. We found that a model combining simplicity theory with CPT came close to matching machine learning algorithms in making outsample predictions. This model uncovered three heterogeneous groups in the population, two complexity averse and one complexity loving. While demographic characteristics do not help predict group membership, nancial literacy does. We applied our model to predict real-world prize-linked savings takeup, and showed that unlike previous models it can simultaneously account for di erent phenomena with the same parameter estimates. Our heterogeneous CPT-Simplicity model may lead to a richer understanding of real-world behavior in other areas as well.

References

Abbey, James D., and Margaret G. Meloy. 2017. Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality. Journal of Operations Management, 53: 63 70.

Abdellaoui, Mohammed. 2000. Parameter-Free Elicitation of Utility and Probability Weighting Functions. Managment Science, 46: 1497 1512.

Andersen, Ste en, Glenn W Harrison, and E Elisabet Rutström. 2006. Choice Behavior, Asset Integration and Natural Reference Points. Working Paper.

Andersen, Ste en, Glenn W. Harrison, Morten Igel Lau, and E. Elisabet Rutström. 2006. Elicitation Using Multiple Price List Formats. Experimental Economics, 9: 383 405.

Andreoni, James, and Charles Sprenger. 2011. Uncertainty Equivalents: Testing the Limits of the Independence Axiom. NBER. https://www.nber.org/papers/w17342.

Bernheim, B. Douglas, and Charles Sprenger. 2020. On the Empirical Validity of Cumulative Prospect Theory: Experimental Evidence of Rank-Independent Probability Weighting. Econometrica, 88: 1363 1409.

Bleichrodt, Han, and Jose Luis Pinto. 2000. A Parameter-Free Elicitation of the Probability Weighting Function in Medical Decision Analysis. Management science, 46: 1485 1496.

Bodoh-Creed, Aaron, Jörn Boehnke, and Brent Hickman. 2018. Using Machine Learning to Predict Price Dispersion. Working paper.

Booij, Adam S., Bernard M. S. van Praag, and Gijs van de Kuilen. 2010. A Parametric Analysis of Prospect Theory's Functionals for the General Population. Theory and Decision, 68: 115 148.

Bradley, P.S, K.P Bennett, and A Demiriz. 2000. Constrained K-Means Clustering. Technical Report MSR-TR-2000-65, Microsoft Research, Redmond, WA.

Brandimarte, P. 2014. Handbook in Monte Carlo Simulation: Applications in Financial Engineering, Risk Management, and Economics. Wiley.

Bruhin, Adrian, Thomas Epper, and Helga Fehr-Duda. 2010. Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. Econometrica, 78: 1375 1412.

Camerer, Colin F., and Teck-Hua Ho. 1994. Violations of the Betweenness Axiom and Nonlinearity in Probability. Journal of Risk and Uncertainty, 8: 167 196.

Cerreia-Vioglio, Simone, David Dillenberger, and Pietro Ortoleva. 2015. Cautious Expected Utility and the Certainty E ect. Econometrica, 83: 693 728.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2017. Willingness to Pay and Willingness to Accept are Probably Less Correlated than You Think. NBER. https://www.nber.org/papers/w23954.

Chmielewski, Michael, and Sarah C. Kucker. 2020. An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science, 11: 464 473.

Cole, Shawn, Benjamin Iverson, and Peter Tufano. 2021. Can Gambling Increase Savings? Empirical Evidence on Prize-linked Savings Accounts. Management Science Pre-print

Conlisk, John. 1989. Three Variants on the Allais Example. The American Economic Review, 79: 392 407.

Conte, Anna, John D. Hey, and Peter G. Mo att. 2011. Mixture Models of Choice under Risk. Journal of Econometrics, 162: 79 88.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39: 1 22.

Dillenberger, David. 2010. Preferences for One-Shot Resolution of Uncertainty and Allais-Type Behavior. Econometrica, 78: 1973 2004.

Efron, B., and R.J Tibshirani. 1993. An Introduction to the Bootstrap. Chapman Hall.

Erev, Ido, Eyal Ert, Alvin E. Roth, Ernan Haruvy, Stefan M. Herzog, Robin Hau, Ralph Hertwig, Terrence Stewart, Robert West, and Christian Lebiere. 2010. A Choice Prediction Competition: Choices from Experience and from Description. Journal of Behavioral Decision Making, 23: 15 47.

Erev, Ido, Eyal Ert, Ori Plonsky, Doron Cohen, and Oded Cohen. 2017. From Anomalies to Forecasts: Toward a Descriptive Model of Decisions under Risk, under Ambiguity, and from Experience. Psychological Review, 124: 369 409.

Etchart-Vincent, Nathalie. 2009. Probability Weighting and the `Level' and `Spacing' of Outcomes: An Experimental Study over Losses. Journal of Risk and Uncertainty, 39: 45 63.

Fan, Chinn-Ping. 2002. Allais Paradox in the Small. Journal of Economic Behavior & Organization, 49: 411 421.

Fehr-Duda, Helga, and Thomas Epper. 2012. Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences. Annual Review of Economics, 4: 567 593.

FRED. 2020. South Africa / U.S. Foreign Exchange Rate. Board of Governors of the Federal Reserve System (US). https://fred.stlouisfed.org/series/DEXSFUS.

Freeman, David J., Yoram Halevy, and Terri Kneeland. 2019. Eliciting Risk Preferences Using Choice Lists. Quantitative Economics, 10: 217 237.

Fudenberg, Drew, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. 2020. Measuring the Completeness of Economic Models. Working Paper. http://economics.mit.edu/ les/20972.

Goodman, Aaron, and Indira Puri. 2020. Arbitrage in the Binary Option Market: Distinguishing Behavioral Biases. Working Paper.

Halevy, Yoram, Dotan Persitz, and Lanny Zrill. 2018. Parametric Recoverability of Preferences. Journal of Political Economy, 126: 1558 1593.

Harless, David W., and Colin F. Camerer. 1994. The Predictive Utility of Generalized Expected Utility Theories. Econometrica, 62: 1251 1289.

Harrison, Glenn W. 1994. Expected Utility Theory and the Experimentalists. In Experimental Economics. Springer-Verlag.

Harrison, Glenn W., and E. Elisabet Rutström. 2009. Expected Utility Theory and Prospect Theory: One Wedding and a Decent Funeral. Experimental Economics, 12: 133 158.

Harrison, Glenn W., Steven J. Humphrey, and Arjan Verschoor. 2010. Choice under Uncertainty: Evidence from Ethiopia, India and Uganda. The Economic Journal, 120: 80 104.

Hauser, David J., and Norbert Schwarz. 2016. Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants. Behavior Research Methods, 48: 400 407.

Huck, Ste en, and Georg Weizsacker. 1999. Risk, complexity, and deviations from expected-value maximization: Results of a lottery choice experiment. Journal of Economic Psychology, 20: 699 715.

Huck, Ste en, and Wieland Müller. 2012. Allais for All: Revisiting the Paradox in a Large Representative Sample. Journal of Risk and Uncertainty, 44: 261 293.

Kahneman, Daniel, and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47: 263 292.

Ke, Shaowei, Chen Zhao, Zhaoran Wang, and Sung-Lin Hsieh. 2020. Behavioral Neural Networks. SSRN. https://ssrn.com/abstract=3633548.

Lusardi, Annamaria, and Olivia S. Mitchell. 2014. The Economic Importance of Financial Literacy: Theory and Evidence. Journal of Economic Literature, 52: 5 44.

Lusardi, Annamaria, and Olivia S Mitchelli. 2007. Financial Literacy and Retirement Preparedness: Evidence and Implications for Financial Education. Business Economics, s42: 35 44.

Mo att, Peter G., Stefania Sitzia, and Daniel John Zizzo. 2015. Heterogeneity in Preferences towards Complexity. Journal of Risk and Uncertainty, 51: 147 170.

Mongin, Philippe. 2019. The Allais Paradox: What It Became, What It Really Was, What It Now Suggests to Us. Economics and Philosophy, 35: 423 459.

Neilson, William, and Jill Stowe. 2002. A Further Examination of Cumulative Prospect Theory Parameterizations. Journal of Risk and Uncertainty, 24: 31 46.

Peysakhovich, Alexander, and Je rey Naecker. 2017. Using Methods from Machine Learning to Evaluate Behavioral Models of Choice under Risk and Ambiguity. Journal of Economic Behavior & Organization, 133: 373 384.

Plonsky, Ori, Apel Reut, Ido Erev, Eyal Ert, and Moshe Tennenholtz. 2017. When and How Can Social Scientists Add Value to Data Scientists? A Choice Prediction Competition for Human Decision Making. CPC White Paper. https://cpc-18.com/wp-content/uploads/2018/03/cpc18-white-paper-march-update.pdf.

Plonsky, Ori, Ido Erev, Tamir Hazan, and Moshe Tennenholtz. 2016. Psychological Forest: Predicting Human Behavior. SSRN. https://ssrn.com/abstract=2816450.

Puri, Indira. 2020. Preference for Simplicity. SSRN. https://ssrn.com/abstract=3253494.

Sonsino, Doron, Uri Benzion, and Galit Mador. 2002. The Complexity E ects on Choice with Uncertainty Experimental Evidence. The Economic Journal, 112: 936 965.

Tanaka, Tomomi, Colin F Camerer, and Quang Nguyen. 2010. Risk and Time Preferences: Linking Experimental and Household Survey Data from Vietnam. American Economic Review, 100: 557 571.

Tversky, Amos, and Craig R Fox. 1995. Weighing Risk and Uncertainty. Psychological Review, 102: 269.

Tversky, Amos, and Daniel Kahneman. 1992. Advances in Prospect Theory: Cumulative Representation of Uncertainty. Journal of Risk and uncertainty, 5: 297 323.

Wu, George, and Richard Gonzalez. 1996. Curvature of the Probability Weighting Function. Management Science, 42: 1676 1690.

Xiao, Jing J., ed. 2008. Handbook of Consumer Finance Research. Springer.

Appendix A Data

A.1 Random Lottery Generation Procedure

There are 50 lotteries, with 10 for each outcome in $\{2, 3, 4, 5, 6\}$. Each random lottery is generated as follows:

1. Pick a range $r \in \{2, 2.5, 3, 3.5, 4\}$.[28]
2. Pick a payment amount x_1 uniformly at random from \$1 - \$10 and such that x_1 is a multiple of \$0.25.
3. Set the second payment amount x_2 to be $\max(x_1 - r, 1)$ if $x \geq 5.5$ and $\min(x_1 + r, 10)$ if $x < 5.5$.
4. Generate outcomes $x_3, \ldots, x_n$ uniformly at random such that each is between x_1 and x_2 and is a multiple of \$0.25.
5. Generate probabilities as follows:
   (a) Step 1: Pick randomly a probability $p_1 \in (0, 1)$ which will be the probability of x_1.
   (b) Step $i < n$: Pick randomly a probability $p_i \in (0, 1 - \sum_{j=1}^{i-1} p_j)$
   (c) Step n: Set $p_n = 1 - \sum_{j=1}^{n-1} p_j$

The resulting lotteries do not differ significantly in mean, variance, or skewness by number of outcomes. Figure 4 plots each measure against the number of outcomes. Table 5 provides the correlation of each measure with the number of outcomes.

	Mean	Variance	Skewness
N. Outcomes	-0.19	-0.12	0.15
	(0.19)	(0.41)	(0.31)

Table 5: Correlation between moments of the lottery and number of outcomes. The correlation coefficient between the number of outcomes and each of mean, variance, and skewness are shown. The p-values are in parentheses.

A.2 Financial Literacy Questions

The financial literacy questions asked are:

1. If the chance of getting a disease is 10 percent, how many people out of 1,000 would be expected to get the disease?

[28] Two of each range were used for each support size.

(a) (b)

(c)

Figure 4: Moments of the Lottery and Number of Outcomes

2. If 5 people all have the winning number in the lottery and the prize is 2 million dollars, how much will each of them get?

3. Let's say you have 200 dollars in a savings account. The account earns 10 percent interest per year. How much would you have in the account at the end of two years?

A.3 Demographic Summary

Table 6: Demographics and Financial Literacy, Overall Sample

Age	
20-29	24.5%
30-39	47%
40-49	18%
50-59	7.5%
> 60	3%
Gender	
Male	64.5%
Female	34%
Prefer not to answer	1.5%
Race	
White	82.5%
Black	11%
Other	6.5%
Education	
High school graduate	11%
Some college	19.5%
2-Year degree	12%
4-Year degree	41.5%
Master's degree	14%
Doctoral degree	1%
Professional degree (ex. JD, MD)	1%
Employment	
Paid employee	65%
Self-employed	18%
Not working	16%
Prefer not to answer	1%
Household Income	
< $10,000	3%
$10,000-$24,999	9.5%
$25,000-$49,999	35%
$50,000-$74,999	24.5%
$75,000-$99,999	12%
$100,000-$124,999	4%
$125,000-$149,999	4.5%
> $150,000	6%
Prefer not to answer	1.5%
Financial Literacy	
Score= 0	2%
Score= 1	29.5%
Score= 2	68.5%
Respondents	200

A.4 Propensity Scores

(a) Complexity Averse, Less Behavioral (b) Complexity Averse, More Behavioral

(c) Complexity Loving

Figure 5: Propensity Scores by Group

Appendix B Estimation

B.1 Finite Data and The Attractiveness of Alternative

Here we show by simulation that with few observations, the log likelihood of incorrect alternative 's can be close to the log likelihood of the parameters used to generate the data. For each model of EU, CPT, and CPT-Simplicity, we generated the data using the parameters listed in Tables 7 - 9, with simplicity parameters $= 2;\ = 2;\ = 1$. For each of n individuals, we generated their certainty equivalent for a given lottery as the CE predicted by the relevant model, plus noise $\sim N(0; 4)$.

For the EU model and CPT models, for each in $A = f0:1; 0:2; 0:3; 0:4; 0:5; 0:6; 0:7; 0:8; 0:9g$, we calculate the log likelihood using the true (in the EU case, $= 1$). For the CPT-Simplicity model, for each $2 A$, we calculate the log likelihood using each of $2 f 0:1; 0:3; 0:5; 0:7; 1:0; 1:5; 2:0g$, $2 f 0:5; 1:0; 2:0g$, and the true . The log likelihood corresponding to a given is defined as the highest log likelihood using that for the same class of model.

Tables 7 - 9 show that, for a wide range of and , regardless of the model used, as the number of individuals increases, the number of s whose log likelihood is close to the global log likelihood, decreases. This demonstrates that alternative s can be relatively more attractive with less data.

Table 7: CPT-Simplicity Model: N. Observations Versus Log Likelihood Distinctiveness

N. Individuals	True	True	's whose log-likelihoods are +/- 10 of true parameter log-likelihood
1	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7
30	0.7	0.8	0.4, 0.5, 0.6, 0.7
100	0.7	0.8	0.6, 0.7
1	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
30	0.7	0.5	0.3, 0.4, 0.5, 0.7
100	0.7	0.5	0.6, 0.7
1	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8
30	0.7	0.3	0.2, 0.3, 0.4, 0.5, 0.6, 0.7
100	0.7	0.3	0.6, 0.7

Table 8: EU Model: N. Observations Versus Log Likelihood Distinctiveness

N. Individuals	True	's whose log-likelihoods are +/- 10 of true parameter log-likelihood
1	0.9	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.9	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.9	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.9	0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
1	0.7	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.7	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.7	0.1, 0.2, 0.3, 0.4, 0.5,0.6, 0.7, 0.8, 0.9
1	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.5	0.1,0.2, 0.3, 0.4, 0.5, 0.6,0.7, 0.8, 0.9

Table 9: CPT Model

N. Individuals	True	True	's whose log-likelihoods are +/- 10 of true parameter log-likelihood
1	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.7	0.8	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
1	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.7	0.5	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
1	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
5	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
30	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
100	0.7	0.3	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9

B.2 Best Initialization

By best initialization, we mean the initialization with lowest validation MSE among those that result in well-de ned train and validation set numbers. In particular, if an initialization results in a non-computable (e.g. NaN) value, we drop it from consideration, as for example if an initialization results in the algorithm taking the square root of a negative number, this would be considered a non-computable value. Dropping initializations with non-computable values is standard practice, see for example the code of Bruhin, Epper and Fehr-Duda (2010).

Appendix C Group Membership

Table 10: Group Membership Regressions

	Complexity Averse Less Behavioral	Complexity Averse More Behavioral	Complexity Loving
Financial Literacy	0.089	-0.125	0.037
	(0.057)	(0.053)	(0.056)
Age	0.024	0.009	-0.032
	(0.032)	(0.027)	(0.030)
Income	0.012	-0.023	0.011
	(0.018)	(0.013)	(0.016)
College-Educated	-0.016	0.091	-0.076
	(0.061)	(0.048)	(0.060)
Male	0.028	0.017	-0.045
	(0.63)	(0.051)	(0.062)
White	-0.081	-0.018	0.098
	(0.071)	(0.060)	(0.065)
Employed	0.007	-0.043	0.032
	(0.081)	(0.064)	(0.077)
N	19600	19600	19600
R^2	0.025	0.052	0.022

Standard errors clustered at the individual level. Standard errors in parentheses.
$p < 0:10$, $p < 0:05$, , $p < 0:01$

Online Appendix I Analysis

I.I Validation Performance

Table 11: Validation Performance of Models by Number of Groups

Model	Groups	Validation MSE
Expected Utility	1	82.04
		(16.13)
	2	79.71
		(15.79)
	3	79.26
		(15.83)
Prospect Theory	1	82.24
		(15.85)
	2	74.24
		(15.06)
	3	72.24
		(14.17)
Simplicity Theory	1	82.38
		(16.01)
	2	75.19
		(15.33)
	3	70.39
		(14.08)
Cumulative Prospect Theory	1	77.82
		(15.19)
	2	70.25
		(13.45)
	3	68.05
		(13.04)
PT-Simplicity	1	82.33
		(2.47)
	2	75.19
		(11.95)
	3	69.96
		(14.26)
CPT-Simplicity	1	77.17
		(15.31)
	2	70.30
		(14.44)
	3	64.31
		(13.12)

Standard deviation in parentheses.

Table 12: Machine Learning Algorithms: Best Validation Performance

Model	Validation MSE
Gradient Boosting Tree	101.2
Neural Network	88.13
K-Means	60.41

I.II Data Visualization: Figures

(a) Expected Value

(b) Variance

(c) Probability Weighting

Figure 6: Data Overall: Non-Parametric Behavior

(a) Expected Value

(b) Variance

(c) Probability Weighting

Figure 7: Group `Complexity Averse, Less Behavioral': Non-Parametric Behavior

(a) Expected Value

(b) Variance

(c) Probability Weighting

Figure 8: Group `Complexity Averse, More Behavioral': Non-Parametric Behavior

(a) Expected Value

(b) Variance

(c) Probability Weighting

Figure 9: Group `Complexity Loving': Non-Parametric Behavior

I.III Data Visualization: Complexity Loving Group

Figure 10 shows that, xing the number of outcomes, the median risk premium is increasing in the variance of the lotteries, e.g. the complexity loving group is also risk averse, without parametric assumptions. The risk premiums are negative because this group is complexity loving.

I.IV CPT Parameter Estimates

Table 13: CPT Group Parameters

	Group 1	Group 2	Group 3
	1.696	0.791	0.790
	(0.019)	(0.031)	(0.027)
	0.756	0.907	0.377
	(0.067)	(0.140)	(0.014)
Number	99.940	64.760	31.300
	(8.080)	(7.486)	(3.371)

Standard deviation in parentheses.

Online Appendix II Varying by number of outcomes

Here we allow to vary by number of outcomes, using the group assignments found by the three-group CPT model, where we assign a person to a group if their propensity score for that group exceeds 60%. Since there are 10 lotteries per number of outcomes, the 80% train, 10% validation split used in the main body of the paper would allocate only 1 lottery to the validation set; we therefore use 80% train, 20% test split here. Within the 80% train data, we use the the Bruhin, Epper and Fehr-Duda (2010) method of taking an average of parameter values over initializations. Figure 11 shows point estimates and 95% con dence intervals for each estimate. For no group are the 's monotonic in number of outcomes. Further, the median test MSE across groups and outcomes is 0.67, worse than the test performance of CPT-Simplicity. [29]

[29] The di erent estimation procedures means that these numbers are not directly comparable. However, because the CPT-Simplicity model was estimated on all lotteries at once, and because it had a lesser percentage of data in the test set, a priori one may have expected it to do worse on the test set than the various tests.

(a) Two Outcomes

(b) Three Outcomes

(c) Four Outcomes

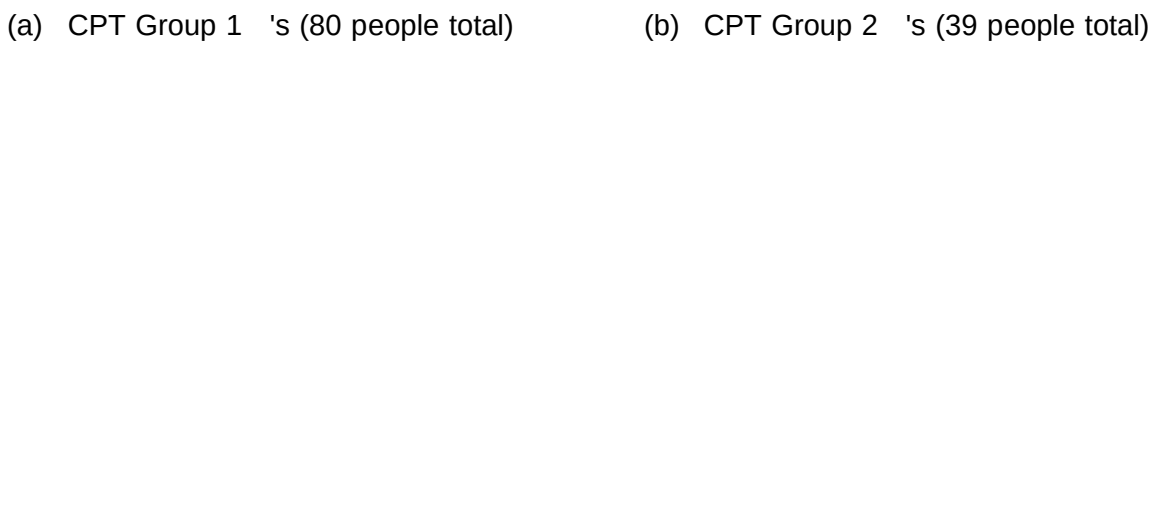
(d) Five Outcomes

(e) Six Outcomes

Figure 10: Risk Premium Versus Variance for Fixed Number of Outcomes: Complexity Loving Group

Online Appendix III Robustness Checks

We check whether alternative forms of data cleaning change our results. Answering 0 (the lowest possible certainty equivalent) many times may reflect a strong preference for simplicity,



(a) CPT Group 1 's (80 people total) (b) CPT Group 2 's (39 people total)

(c) CPT Group 3 's (27 people total)

Figure 11: Allowing to Vary by Number of Outcomes

but it could also re ect a lack of attention to the survey. Similarly, answering 10 (the highest possibly certainty equivalent) many times may re ect an extreme preference for risk, but it could also re ect a lack of attention. (Figure 12 plots the number of times an individual chooses 0 or 10 against the amount of time they take to complete the survey. We next check whether there are some individuals who always answer 0 or 10 (Figure 13). Finally, we check whether there are any lotteries for which respondents put `0' or 10 disproportionately often (Figure 14). There are no strong anomalies in the data. There are four individuals who always chose 10 as their responses and no individuals who always chose 0. Eight out of our 200 individuals choose 0 or 10 for 15 or more of their 25 responses. The completion time for individuals is slightly less for individuals with more 0s or 10s, but not in a way that is statistically signi cant. In lotteries, there are some which receive 30 (out of around 100) responses of 10, but this happens for several lotteries and not one in particular. No lottery receives more than 6 (out of around 100) responses of 0.

Figure 12: Survey Completion Time Versus Number of 0s or 10s

We show that the results remain qualitatively unchanged if we drop no data, if we drop the four individuals who always chose 10 (this was used in the body of the paper), or if we drop the eight individuals who chose 0 or 10 at least fteen times. For all these forms of data cleaning, the CPT-Simplicity model is close to machine learning performance, and nds three groups: two complexity averse, with one distorting probabilities more than the other, and one complexity loving.

Figure 13: Number of 0s or 10s by Individual

(b) Number of 0s

(d) Number of 10s

Figure 14: Number of 0s or 10s by Lottery

(b) Number of 0s

(d) Number of 10s

III.I Results, No Drops

Table 14: Test Performance, No Drops

Model	Test MSE	ML Completeness Score
Expected Value	88.02	0%
EU	85.60 (12.63)	11.80%
PT	76.72 (11.38)	55.10%
Simplicity	75.19 (11.54)	62.55%
PT-Simplicity	74.91 (11.06)	63.92%
CPT	74.00 (10.65)	68.36%
CPT-Simplicity	69.37 (10.76)	90.93%
Machine Learning Benchmark	67.51 (8.15)	100%

Standard deviation in parentheses.

III.II Results After Dropping People With Over 15 Responses of 0 or 10

There are eight people (4% of the total sample) who give the highest or lowest certainty equivalent in response to more than 15 of 25 of their lottery questions. As a robustness check, we drop these people and report results here.

Table 15: Test Performance, Dropping People With Over 15 0's or 10's

Model	Test MSE	ML Completeness Score
Expected Value	75.18	0%
EU	76.75 (12.09)	-11.15%
PT	69.67 (10.89)	39.13%
Simplicity	67.25 (11.21)	56.32%
PT-Simplicity	66.61 (10.40)	60.87%
CPT	65.07 (9.93)	71.80%
CPT-Simplicity	61.45 (10.02)	97.51%
Machine Learning Benchmark	61.10 (7.08)	100%

Standard deviation in parentheses.

III.III Alternative Bounds

Prior experimental work (Fehr-Duda and Epper (2012), Bruhin, Epper and Fehr-Duda (2010), Kahneman and Tversky (1979)) typically finds the risk aversion parameter to be in the range (0.8,1.0). The estimation procedure used for the results in the main body of the paper restrict to lie between 0.7 and 1.8. These are also the bounds that result when we estimate a three-group CPT model on lotteries with two outcomes only, in the spirit of Bruhin, Epper and Fehr-Duda (2010). The lowest less one standard deviation, and the highest plus one standard deviation, give these bounds. For completeness, Table 16 reports results when the bounds are made more flexible, at between 0.5 and 2.0. The ranking of preference models is similar to that reported in the main body of the paper.[30]

[30] However, when loosening the bounds, we obtain parametric results which are at odds with the non-parametric group by group analysis.

Table 16: Test Performance, Alternative Bounds

Model	Test MSE	ML Completeness Score
Expected Value	82.87 (12.26)	0%
EU	77.94 (11.61)	29.59%
PT	73.56 (10.94)	52.21%
Simplicity	72.74 (8.42)	56.69%
PT-Simplicity	69.96 (10.17)	72.24%
CPT	68.65 (10.30)	79.57%
CPT-Simplicity	65.54 (10.37)	96.98%
Machine Learning Benchmark	65.00 (6.88)	100%

Standard deviation in parentheses.

Online Appendix IV Applications

IV.I Event Splitting

Figures 15 and 16 show the results of our simulations for the split low and split high tests of Bernheim and Sprenger (2020), respectively. Observe that the disutility incurred from splitting is over and above the CPT prediction for the complexity averse groups, while the utility added from splitting is over and above the CPT prediction for the complexity loving groups.

Online Appendix V Experimental Instructions

Screenshots of the experimental instructions and comprehension questions are shown in Figures 17 - 23.

There were five attention check questions scattered throughout the survey:

What is the last name of a man named Richard Smith?

ˆ Richard

Figure 15: Split Low

Figure 16: Split High

ˆ Bob

ˆ Smith

Which of the following is a color?

ˆ artisan

ˆ sentry

ˆ festival

ˆ orange

An omelette is made of...

ˆ circle

ˆ oval

ˆ eggs

ˆ pan

Ed Sheeran is a...

ˆ bird

ˆ human

ˆ ying squirrel

Through which website are you taking this survey?

ˆ Neopets

ˆ Pokemon

ˆ mTurk

ˆ Teletubbies

Finally, Table 17 lists the 50 randomly generated lotteries used.A participant was asked to provide certainty equivalents either for the rst set of 25 or the second set of 25, with this assignment made at random.

Table 17: Lotteries

Lottery Number	O1	P1	O2	P2	O3	P3	O4	P4	O5	P5	O6	P6
1	8	0.53	6	0.47								
2	6	0.69	3.5	0.31								
3	7	0.56	4	0.44								
4	8.75	0.28	5.25	0.72								
5	8.25	0.11	4.25	0.89								
6	5	0.32	4.25	0.41	3	0.27						
7	9.75	0.55	9	0.22	7.25	0.23						
8	7	0.16	6.25	0.29	4	0.55						
9	8.25	0.01	7.75	0.13	4.75	0.86						
10	10	0.82	7	0.12	6	0.06						
11	10	0.78	9	0.02	8.25	0.08	8	0.12				
12	6.75	0.04	6.25	0.07	4.5	0.02	4.25	0.87				
13	6.75	0.77	6.5	0.13	5.75	0.02	3.75	0.08				
14	8	0.26	7.25	0.14	7	0.46	4.5	0.14				
15	8	0.19	7.75	0.09	6	0.48	4	0.24				
16	7.75	0.6	7.5	0.34	6.75	0.04	6.5	0.01	5.75	0.01		
17	7.5	0.03	7.25	0.21	6.75	0.06	5.25	0.01	5	0.69		
18	8.75	0.63	8	0.06	6.75	0.02	6.25	0.05	5.75	0.24		
19	7	0.18	5	0.01	4.75	0.03	4.5	0.07	3.5	0.71		
20	5	0.12	4.5	0.04	2.75	0.03	1.25	0.54	1	0.27		
21	6	0.48	5.75	0.01	5.5	0.05	5	0.09	4.25	0.19	4	0.18
22	5.25	0.46	5	0.03	4.75	0.18	3.5	0.01	3.25	0.02	2.75	0.3
23	9.5	0.01	8.75	0.01	8.5	0.03	8	0.01	6.75	0.01	6.5	0.93
24	8.25	0.15	8	0.24	7.75	0.06	7.5	0.02	7	0.03	4.75	0.5
25	9.5	0.66	9.25	0.2	8.75	0.06	8.5	0.01	6.5	0.04	5.5	0.03
26	6.5	0.87	4.5	0.13								
27	8.25	0.81	5.75	0.19								
28	6	0.79	3	0.21								
29	8	0.21	4.5	0.79								
30	6.5	0.51	2.5	0.49								
31	6.5	0.01	6	0.54	4.5	0.45						
32	8.5	0.1	6.25	0.17	6	0.73						
33	6.75	0.15	6	0.08	3.75	0.77						
34	9.5	0.26	6.5	0.24	6	0.5						
35	8.25	0.5	4.5	0.45	4.25	0.05						
36	6.5	0.28	6.25	0.01	6	0.08	4.5	0.63				
37	4.75	0.35	4.25	0.42	3.75	0.14	2.25	0.09				
38	4.75	0.19	3.75	0.07	3.5	0.18	1.75	0.56				
39	9.25	0.01	8.25	0.68	7.25	0.18	5.75	0.13				
40	9.5	0.16	7	0.41	5.75	0.1	5.5	0.33				
41	4.5	0.2	4	0.32	3.75	0.01	3.25	0.04	2.5	0.43		
42	5	0.56	4.5	0.19	3.5	0.03	3.25	0.03	2.5	0.19		
43	6.25	0.64	5.25	0.06	5	0.02	4.5	0.18	3.25	0.1		
44	6.5	0.1	6.25	0.07	4.25	0.14	3.75	0.07	3	0.62		
45	6.75	0.3	6.5	0.06	5.25	0.01	4.75	0.1	2.75	0.53		
46	4	0.02	3.75	0.02	3	0.01	2.75	0.09	2.5	0.08	2	0.78
47	7.25	0.02	7	0.04	6	0.01	5.5	0.03	5	0.25	4.75	0.65
48	4.75	0.51	4.25	0.06	3.5	0.03	2.25	0.01	2	0.05	1.75	0.34
49	10	0.49	8	0.06	7.5	0.04	7	0.05	6.75	0.04	6.5	0.32
50	5	0.21	4	0.01	3.25	0.01	3	0.01	1.25	0.02	1	0.74

Figure 17: Experimental Instructions

Figure 18: Experimental Instructions Continued: Sample Question 1

Figure 19: Experimental Instructions Continued: Sample Question 2

Figure 20: Experimental Instructions Continued: Sample Question 3

Figure 21: Experimental Instructions Continued: Comprehension Question 1

COMPREHENSION QUESTION

Suppose Anna responds to the below question as follows.

For the purpose of this question, LOTTERY L is:

Probability	Outcome
50.0%	$6.0
50.0%	$8.0

Consider the following questions:

	Option S	Option R
Q1.	$5.50 for sure	Lottery L
Q2.	$5.78 for sure	Lottery L
Q3.	$6.06 for sure	Lottery L
Q4.	$6.34 for sure	Lottery L
Q5.	$6.62 for sure	Lottery L
Q6.	$6.90 for sure	Lottery L
Q7.	$7.18 for sure	Lottery L
Q8.	$7.46 for sure	Lottery L
Q9.	$7.74 for sure	Lottery L
Q10.	$8.02 for sure	Lottery L

I choose Option R for Questions 1 - __, I choose Option S for Questions (__ + 1) - 10.

- 0
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

Which of the following is true?

- Anna prefers $5.50 for sure to Lottery L
- Anna prefers $8.02 for sure to Lottery L
- Anna prefers Lottery L to $6.62 for sure

Figure 22: Experimental Instructions Continued: Comprehension Question 2