Evaluating models of choice under risk and ambiguity using methods from machine learning

Alexander Peysakhovich, Yale University*[1]
Jeffrey Naecker, Wesleyan University[2]

Abstract

How can we incorporate machine learning into understanding human choice under uncertainty? We use an online platform to collect a large data set of individuals indicating their willingness to pay to play uncertain lotteries. We then calibrate both formal models and machine learning methods on a subset of individuals' decisions and attempt to predict future decisions by the same individuals. We show that economic models tend to do worse out-of-sample than would be expected from their in-sample performance, so we argue that "in-sample variance explained" estimates should be taken with caution. In addition, models that attempt to fit a single risk profile to a large population do very poorly. In the domain of risk, a version of expected utility that allows for non-linear probability weighting (as in cumulative prospect theory) and individual-level parameters performs as well out-of-sample as machine learning techniques. In the domain of ambiguity, however, two of the most widely studied models (a linear version of maximin preferences and second order expected utility) fail to compete with the machine learning methods. These results suggest that there are potentially higher returns from working on a portable ambiguity aversion model than on improving existing models of decisions under risk. Our results highlight ways in which behavioral scientists can incorporate machine learning techniques in their daily practice to gain genuinely new insights. They also underscore the importance of looking out-of-sample when evaluating model quality.

[1] alex.peys@gmail.com Current author affiliation: Core Data Science Team, Facebook Inc. These studies were performed and most of the analysis was done while the author was a Research Scientist at the Human Cooperation Lab at Yale University.
[2] jnaecker@wesleyan.edu
* - Corresponding author. We thank Dean Eckles, Ido Erev, Drew Fudenberg, Alex Imas, Jon Levin, Muriel Niederle, David Rand, Alvin Roth, Andrei Schleifer, Sean Taylor, Erez Yoeli and participants at the IC2S2 Conference as well as the Stanford Experimental Economics course for valuable comments and suggestions. Errors remain our own. Peysakhovich thanks the John Templeton Foundation for financial support.

Decisions ranging from the mundane (e.g. choosing a restaurant) to the life-changing (e.g. choosing a job) include elements of uncertainty. For this reason, understanding how individuals evaluate uncertain prospects has been a key research area in the behavioral and social sciences for nearly two centuries (Bernoulli 1738, Kreps 1988).[3] This has led to the creation of simple mathematical models that are characterized by parameters with intuitively understandable interpretations (e.g. the coefficient of risk aversion). There are many important recurring questions in any research program: How good are these models? What commonly used assumptions are the most restrictive? What domains of uncertainty appear to be potentially fruitful targets for theorists?

In this paper we combine techniques from the literature on machine learning with standard behavioral science. Our aim is to show how these methods can help shed light on these questions. We focus on two domains: risk (Camerer 1995, Kahneman & Tversky 2000, Kreps 1988, Savage 1954), where the probability of an uncertain outcome is perfectly known, and ambiguity (Knight 1921, Ellsberg 1961, Camerer & Weber 1992, Trautmann & Van De Kuilen 2013), where decision-makers have partial, but not full, information to estimate the likelihood of an outcome. We recruit over 600 participants to indicate their willingness to pay for uncertain prospects whose features are randomly generated. As is common in the statistical learning literature (Friedman et al. 2009), we take a subset of these individuals' decisions as an out-of-sample "test set." We calibrate on the remaining "training set" several economic models: expected utility (EU, von Neumann & Morgenstern 1945) and expected utility with non-linear probability weighting (EUP, Tversky & Kahneman 1992, Prelec 1998) in the case of risk and second-order expected utility (SOEU, Grant et al 2009) and maximin preferences (MM, Gilboa & Schmeidler 1989, Levy et al. 2010, Tymula et al. 2012) in the case of ambiguity on the remaining decisions. We then ask: how well do these models predict the held out test set decisions?

This exercise allows us to tackle two issues: first, it allows us to consider the relative explanatory power of the economic models. Note that because EUP nests EU but has an additional parameter, it will always fit (weakly) better in-sample. However, this may simply be over-fitting and the more complicated model may actually do worse out-of-sample. Thus comparing the models' out-of-sample fit allows us to ask whether the additional model complexity adds value in terms of playing an important part in explaining variation in behavior in problems the model has not yet

[3] This should not be confused with the discipline of statistics/decision science which is generally concerned with how individuals *should* evaluate uncertain prospects (Savage 1954).

encountered.[4] This focus on out-of-sample error is justified strongly by our data: we find that in-sample errors are substantially optimistic estimates of out-of-sample error for economic models.

Of course, a statement that a model explains X% of the variance in a particular domain begs the question: is that good or bad? A model that predicts 10% of the variance in a very clean data set might be considered to have quite poor explanatory power. However, if there is substantial noise (either due to sampling error, poor data construction, or other factors), explaining 10% of the variance may actually be quite good.

Thus, we are interested in *explained variance* as a proportion of *explainable variance*. To estimate explainable variance, we turn to tools from machine learning (ML). These tools are designed specifically for prediction and so we use their accuracy on the test set as an estimate of explainable variance in our experiments. As our ML benchmark model we use a cross-validated regularized regression. To allow linear regression to fit non-linear functions we take a basis expansion of all potential decision-relevant variables (probabilities and prizes for each outcome) as well as their interactions. In our most powerful model we also include interactions of each decision-relevant variable with subject-level dummies. This gives us 55,000+ parameters to estimate, so to prevent overfitting we cross-validate and regularize the model (i.e. penalize the model for complexity).

We find that the regularized regression outperforms expected utility models by a large margin. We also find that attempting to fit representative agent models without allowing for individual-level heterogeneity makes the predictive power of any model not much better than the simple prediction given by the empirical average willingness to pay over all prospects in the data set. However, when individual level parameters are allowed EUP does as well as the machine learning algorithms. We interpret this as a victory for probability weighting: this parameter increases out of sample prediction considerably, so it is an important feature of models of uncertain choice. We also consider this a victory for the economic models: a ~600 parameter model (2 per person x ~300 subjects) that is interpretable (i.e. the coefficient of risk aversion has an economic meaning outside of the model) is able to predict choices as well as the ML algorithm

[4] This out-of-sample comparison is quite common in the machine learning literature but surprisingly rare in the social science literature. Though some experimental economists have focused on predictive power of models (Erev et al. 2010, Camerer 2003) and data mining and machine learning approaches are beginning to appear in more experimental work (Fudenberg & Peysakhovich 2014, Naecker 2014) finance (Moritz & Zimmerman 2014), time series analysis (Varian 2014) and political science (Grimmer 2015).

which has two orders of magnitude more parameters (~55,000) and is optimized purely for prediction and not interpretability.

On the other hand, in the domain of ambiguity we find that neither second order expected utility nor maximin preferences are able to predict individual out-of-sample choices as well as the ML models.. We interpret this as an opportunity for empirically-minded theorists: these results, combined with the success of the EUP in the domain of risk, suggest there is ample room for the development of a simple model for the domain of ambiguity that predicts well and yet is relatively parsimonious.

Choice Under Risk

Experimental Design

Our first experiment focuses on the domain of risk. Participants were recruited from Amazon Mechanical Turk and were compensated for their time with rates standard in the literature. All research was approved by the Institutional Review Board of Harvard University.

All decisions made were hypothetical but participants were instructed to treat each decision as if it were real. While online experiments are much less controlled, faster and have smaller stakes than traditional lab sessions there is substantial evidence that standard behavioral economic effects replicate on Mechanical Turk (Peysakhovich & Rand 2015, Imas 2014, Fudenberg & Peysakhovich 2014, Naecker 2014), the pool is more representative (Paolacci & Chandler 2014) and that the size of stakes (even the use of pure hypotheticals) matters little (Amir & Rand 2012, Peysakhovich & Karmarkar 2015). There are known issues with Mechanical Turk samples, for example that participants are well experienced with experimental paradigms, much more so than student populations (Rand et al. 2014). Though we acknowledge this potential confound in our context it is more likely a feature rather than a bug – more "professional" participants are more likely to understand the task at hand and if anything are more likely to have stable, measurable preferences rather than noise due to confusion or unfamiliarity with the task.

In our experiments participants were faced with choices about *lotteries*. Lotteries were described as being an urn containing 100 balls, some of which were red, some of which were blue and some of which were green. Each color had an associated monetary prize.

Participants were asked to enter their willingness to pay (WTP) to *play* each lottery. A lottery was played out as follows: A ball would be drawn randomly from the urn and the participant won the

amount of money associated with the ball. Each lottery was presented as tables like the one below.

	Red	Blue	Green
# Balls	25	14	61
Prize	$10	$2	$0

Participants were educated on how to read the tables as well as the rules of the game. Participants completed a comprehension quiz before starting the experiment; we remove data from individuals who answered this quiz incorrectly as well as those who do not finish the full experiment ($N_{recruited}$ = 350, N_{sample} = 315). See the online appendix for full experimental instructions.

For each experiment we randomly generated large sets of potential lotteries by randomizing the features $\{p_{red}, p_{blue}, p_{green}, money_{red}, money_{blue}\}$ with $money_{green}$ always being 0. Probabilities were generated uniformly at random (subject to the constraint that they sum to unity) and prizes were generated from the uniform distribution from $5 to $30. Participants entered their WTP for 10 such randomly generated lotteries.

We split the data into a randomly selected *training set* of 7 questions per individual and *test set* of 3 questions per individual. Our core analysis involves using the training set to calibrate different models of individual decision-making and then use the test set to see how well these models can do at predicting choices they have not seen before. We use *mean squared error* as our metric.

Models

For choices under risk we consider expected utility as our baseline model. In particular, we choose exponential expected utility (sometimes called constant absolute risk aversion or CARA utility Mas-Collel et al. 1995). Formally we model that the utility of a lottery is given by:

$$EU(L) = p_{red}\,(money_{red})^{\alpha} + p_{blue}\,(money_{blue})^{\alpha}$$

Where α is the coefficient of risk aversion with 1 being risk neutrality and 0 being complete risk aversion. Of course one then needs to transform this into a WTP for that lottery. We incorporate this into our model by assuming that the stated WTP of individuals for the lottery is their

certainty equivalent, that is, the utility of this amount of money is equal to the utility of the lottery. Thus, we derive the equality

$$WTP^{\alpha} = p_{red}\,(money_{red})^{\alpha} + p_{blue}\,(money_{blue})^{\alpha}.$$

The EU model assumes that probabilities enter into utility linearly. However, there is substantial evidence that this is not the case: individuals appear to overweight small probabilities, behaving as if they are larger than they actually are, and underweight large probabilities (Kahneman & Tversky 1979, Tversky & Kahneman 1992). To incorporate this into a more flexible expected tility with probability weighting (EUP) model we use the functional form explored by Prelec (1998):

$$EU(L) = g(p_{red})(money_{red})^{\alpha} + g(p_{blue})(money_{blue})^{\alpha}$$

Where the probability weighting function *g(p)* is given by

$$g(p) = \frac{p^{\gamma}}{\left(p^{\gamma} + (1-p)^{\gamma}\right)^{1/\gamma}}$$

This gives us a second parameter, γ, which characterizes an individual's probability weighting function with γ=1 returning the standard linear weighting.

We consider two classes of models: in one, we fit a single parameter (or pair of parameters in the case of EUP) for the full population of individuals (i.e. what an economist would refer to as a *representative agent* model). In the more complex class of models, we allow for individual-level risk (and probability weighting) parameters. Going from a representative agent to heterogeneous agents increases the model's number of parameters by an order of 300, so we view it as important to see the improvement in predictive accuracy from relaxing the representative agent assumption. In all our estimates we allow risk aversion and probability weighting parameters to range between 0 and 1. Thus we require individuals to be risk averse and do not allow them to overweight probabilities in a "strange" manner. Relaxing this constraint to allow [0,2] for both values only decreases out of sample fit.

Before we turn to showing the predictive power of economic models, we discuss our benchmark ML models.

Machine Learning Methods

As our benchmark we use regularized regression. We give a brief overview of the procedure here and direct the reader to more specialized texts (eg. Freidman et al. 2009) for more discussion on the derivation and Bayesian interpretation of the regularized regression technique as well as a more in-depth discussion of cross-validation and the bias-variance tradeoff.

The regularized regression optimization problem is similar to the one use in OLS estimation, however it incorporates a penalty for model complexity. This means that we get a model whose coefficients are biased (ie. $E(b_i)$ is not the true value beta), however the error that this bias introduces is, theoretically and in practice, offset by the fact that the regularized model does not "chase noise" and overfit in sample, and so we get a lower mean-squared error overall.

Recall that we are using 7 (randomly selected) decisions from each individual to fit the model and the 3 other decisions (never before seen by the model!) to estimate its performance.

The formal objective function is as follows, with ||.|| referring to standard norm notation, y referring to a set of outcomes, **X** being a set of matrix of features (one row for each outcome) and **B** being a vector of regression coefficients:

$$\tfrac{1}{2} ||y - \mathbf{XB}||_2 + \lambda ||\mathbf{B}||_p$$

The idea behind regularized regression is as follows (see Friedman et al. 2009 for a more thorough introduction as well as the mathematical derivation): we run a regression that includes a very large number of features that can be used for predicting the outcome. For this paper each row in the data set is a single decision made by an individual, the feature set is the vector given by p_{red}, p_{blue}, p_{green}, $money_{red}$, $money_{blue}$ as well as quadratic terms for each of these (thus starting with 8 features). In addition we include all up to three-way interactions between these features (resulting in 175 features).[5] This gives us our representative agent model. We use this term because the fact that a particular data point comes from one or another individual is not

[5] There are many potential ways to go from a linear model to represent a more complex set of functional forms. We choose the polynomial basis expansion because it is the simplest to implement.

incorporated into the model at this point. To build the individual level version we interact this feature set with a full set of dummies for each individual (thus giving us an ~55,000 coefficients to be estimated).[6]

We trade off prediction accuracy (in-sample squared residuals) with a penalty. The second term in the above objective function is moderated by the penalty term λ, which is used to prevent overfitting in-sample. A higher λ means that the resulting coefficients will be "shrunk" towards 0 but will also mean that the model will not be as sensitive to the data and thus should be less prone to overfitting and doing poorly in out-of-sample tests. Sending λ to 0 returns the OLS estimates, while setting Lambda to infinity returns a constant model (since we do not penalize the intercept term).

Note that because the penalty is applied to the coefficients, re-scaling the features can change the penalty. The package we use (*glmnet* in R; Friedman et al. 2009) standardizes all features (by transforming them to mean 0, variance 1) when it does the fitting and then unscales them to get the coefficients for the features in their original scales.

Penalizing the coefficient sizes gives us another important advantage: it would be impossible to use standard OLS estimation in this case as the number of columns (>55,000) exceeds our number of rows (7 decisions x 300 subjects = 2100 rows), however because of the added penalty the minimization problem is well specified and easily solvable.

There is also a choice of p in the second term of the objective. We consider the choice of $p=1$ (a linear penalty on coefficient size, commonly called lasso) or $p=2$ (a penalty on the square of each coefficient, commonly called ridge regression). Lasso is a way of introducing sparsity into the model: given a linear penalty, coefficients that add less than a certain amount of predictive power are essentially rounded down to 0. This can be used either because we really believe the true model is sparse (i.e. there are many potential features but only a small number of them actually affect the outcome) or because we want a simpler, more interpretable output of our machine learning procedure (perhaps at the cost of predictive power).

[6] Note that this looks like a large number of parameters, but in fact many entries of our model matrix are 0. This allows us to use sparse matrix methods to efficiently estimate our regressions.

By contrast, p=2 gives us ridge regression which shrinks all estimated coefficients towards 0, but does not provide the sparsification of the lasso. There is a deep connection between both types of penalized regression and Bayesian modeling – in essence it amounts to putting a Normal prior (in the case of ridge) or a Laplace prior (in the case of lasso) on the regression coefficients with lambda playing the role of the prior variance) but this is beyond this basic discussion.[7] We estimate both models on our data but we expect that the ridge should outperform the lasso as there is little reason to expect sparsity in our set of basis expansions.

Note that this means λ is a free parameter. Intuitively, one can think of λ as a shadow price for buying "model complexity," where higher λ pushes us towards a simpler model while lower λ allows us freedom to build more complex functions. How do we then set the optimal price for coefficient? We choose it by cross-validation: we split our training set into sub-sets called folds.

We split the data into 7 folds, each including 1 decision per individual. We then train the model for varying levels of Lambda on 6 folds and predict out to the last one (in essence, we simulate a test-train procedure). We repeat this for each possible "hold out" fold, thus calculating out-of-sample error for each fold. We choose the λ that gives us the smaller average error across these 7 folds. We then put the training set back together and use this chosen lambda as our final penalty parameter.

Figure 1: *We test our models by holding out 3 decisions per individuals calibrating parameter values on the remaining data and asking the models to predict the held out WTP (left panel). To set regularization parameters for our ML models we use cross-validation: in the training set we split the data into 7 folds and for each fold* k *we compare the*

[7] More complex priors can be expressed by changing feature scaling. For example, if we scale one of the inputs to be mean 0 variance 1/100 then adding coefficient (in original units) to this input becomes much "cheaper." This corresponds to placing a prior that has more weight away from 0 on that particular input. We do not employ this method here.

mean-squared errors of models with varying levels of regularization trained on the other folds (right panel shows a single iteration of this procedure).

While this may look complex, at the end of the day, this procedure is simply a way to choose a model from a relatively complicated model space while attempting to combat overfitting in-sample that will lead to bad out-of-sample predictions.

Results

Figure 2 shows our results for risky prospects. For each approach described previously, we plot the mean squared error on both the test set and the training set. We find that the regularized regression outperforms EU by a large margin. We also find that assuming a representative agent (i.e. not allowing for individual-level heterogeneity) makes the predictive power of any model not much better than the simple prediction given by "guess the empirical average willingness to pay over all prospects in the data set" (which yields a mean-squared error of 33).

However, individual-level EUP performs as well as the machine learning algorithms. We interpret this as a victory for probability weighting: this parameter increases out of sample prediction considerably, so it is an important feature of models of uncertain choice. We also consider this a victory for the economic models more generally: a model with 2 parameters per person that is interpretable (i.e. the coefficient of risk aversion has an economic meaning outside of the model) is able to predict choices as well as the ML algorithm which has an order of magnitude more parameters and is optimized purely for prediction and not interpretability. Finally, we find that making a sparsity assumption (ie. using the Lasso) decreases predictive accuracy substantially: this means that in our basis expansion there are many terms which have small coefficients but together contribute substantially to the predictive accuracy of the model; forcing those terms to 0, as the *L1* penalty does, decreases predictive ability by a large factor.

Figure 2: *ML methods outperform standard expected utility, but not expected utility with probability weighting. The representative agent assumption, where all individuals are assumed to have the same utility function, is highly restrictive.*

Choice Under Ambiguity

Experimental Design

Our second experiment focuses on the domain of ambiguity. Participants were again recruited from Amazon Mechanical Turk and were compensated for their time. All decisions made were hypothetical but participants were instructed to treat each decision as if it were real.

Participants again made choices about lotteries. Participants in this experiment did not participate in the risk experiment reported above. This experiment used the procedure introduced in Levy et al. (2010) and further updated in Peysakhovich & Karmarkar (2015). Lotteries were described as being an urn containing 100 balls, some of which were red, some of which were blue. However, unlike in the risk experiment, participants did not know the full composition of the urn. Rather, to induce ambiguity participants lacked all the necessary information to estimate the probability of an outcome.

Participants received the following partial information about each lottery: they knew that there were 100 balls total in the urn. Each ball was colored red or blue. Participants knew that there were *at least X* red balls and *at least Y* blue balls in the urn. However, they did not know the

colors of the remaining 100-*X*-*Y* balls.

Participants were asked to enter their willingness to pay (WTP) to play each lottery. A lottery was played out as follows: A ball would be drawn randomly from the urn and the participant won the amount of money associated with the ball.

Each lottery was presented as tables like the one below.

	Red	Blue	Unknown
# Balls	At least 20	At least 31	49
Prize	$10	$0	??

Participants were education on how to read the tables as well as the rules of the game. Participants completed a comprehension quiz before starting the experiment; as in the risk experiment, we remove data from individuals who answered the comprehension quiz incorrectly as well as those who do not finish the full experiment ($N_{recruited} = 350$, $N_{sample} = 287$). See the online appendix for full experimental instructions.

For each experiment we randomly generated a large set of potential lotteries by randomizing the features $\{X_{red}, Y_{blue}, prize\}$. Probabilities were generated uniformly at random and prizes were generated from the uniform distribution from $5 to $30. Participants entered their WTP for each of 10 such randomly generated lotteries.

Models

We focus on two simple models that have been developed in the ambiguity literature and used in experimental work. The first is a linear version of the maximin model of Gilboa & Schmeidler (1989) that has been used in experimental work such as Peysakhovich & Karmarkar (2015), Tymula et al. (2012), Levy et al. (2010). We follow the exposition in Peysakhovich & Karmarkar (2015) to introduce this model.

Consider a decision-maker facing an ambiguous lottery with a single prize *z*. The mathematical primitives of a lottery are: a set of states of the world, say the interval [0,1], and a winning function g:[0,1] to [0,1]. Given a state of the world *w*, the probability of winning the prize *z* is

given by g(w). The states are ordered such that γ is increasing, namely higher states always mean a (weakly) higher probability of winning the prize. In our case the set of states of the world is given by {0,.01,.02, .03,….,1} which is the set of all possible bag compositions: w=.49 corresponds to the urn having 49 red balls and 51 blue ones, for example The winning function here is then w/100.

The decision-maker does not know w but receives partial knowledge about what it could be: he can conclusively rule out that the state is less than some X and can also rule out that it is greater than Y.

Given this information, the decision-maker builds a probability distribution $p(X,Y)$ on the set of states. We assume that this is done in a Bayesian manner: the decision-maker begins with a full-support prior p_0 on the state space and updates it in accordance with Bayes rule given the knowledge (X,Y) he has. We assume that the prior is uniform on the state space, thus updating with (X,Y) gives a posterior that is uniform on the interval [X, 100-Y].

Let $P(X,Y)$ be the subjective probability of winning the prize given (X,Y). This is simply:

$$P(X,Y) = \int g(w)dp(X,Y)$$

We assume that p is well-behaved so this integral is well defined. The decision maker has the utility function:

$$U(X,Y,z) = (1-\gamma(X+Y))\,P(X,Y)\,z^{\alpha}$$

where γ governs the strength of ambiguity aversion. Note that if γ is zero, the DM acts as an EU agent. Note also in the case of uncertain decisions with no ambiguity (that is, when X + Y = 1) the DM also behaves as an EU maximizer. However, when X+Y is less than 1 the DM "downweights" the probability P(X,Y) by γ and so behaves in an ambiguity averse manner. Here z^{α} is just a standard CARA utility function as in the case of risk.

The maximin model above has an interpretation that the decision-maker shades his beliefs towards the worst possible state of the world. Another, closely related, way to model ambiguity is to assume that individuals treat "objective" uncertainty (e.g. coin flips) differently from

subjective uncertainty (e.g. sports matches between two unknown to the individual teams). This is a key feature of many ambiguity aversion models (Segal 1987, Gul & Pesendorfer 2014, Maccheroni et al. 2006, Klibanoff et al. 2005). We focus on one model of this type, second order expected utility (Grant et al. 2009).

Again, we keep the same setup of states of the world/winning functions/prizes/information as in the exposition above. Except now, we write the utility function as

$$U(X,Y,z) = \int (g(w)\, z^{\alpha})^{\gamma}\, dp(X,Y)$$

Note that setting $\gamma = 1$ gives us back standard expected utility because we simply get $P(X,Y)*z^{\alpha}$. Note also that if p(X,Y) is degenerate (ie. there is no subjective uncertainty about states of the world) then we again get back EU. However, when p(X,Y) is not a point mass then the resulting expected utilities are hit by γ. Thus, an "objective" lottery which is known to have 50-50 odds is preferred to the compound lottery with the average same probability of winning but which includes subjective uncertainty (for example: having odds that could be uniformly drawn from 0-100 to 100-0 but are on average 50-50). This is another way to represent uncertainty aversion that is related to, but not exactly the same as, the maximin model above.[8]

Results

Unlike in the risk domain we find that neither second order expected utility nor maximin preferences are able to predict individual out-of-sample choices as well as ridge regression (Figure 3). We interpret this as an opportunity for experimentally-minded theorists: these results suggest there is ample room for the development of a simple model for the domain of ambiguity that predicts well and yet is relatively parsimonious.

[8] We also point out that this model is much more complicated from a computational standpoint. To compute the utility estimate of an ambiguous lottery one must take a numerical integral with respect to the measure p(X,Y). In our discrete case with 100 states of the world this is not that difficult, however in other situations where the state space is more complicated this model become prohibitive to fit because each iteration of an optimization routine will require the computation of this numerical integral. Computational complexity is not something that decision theorists have generally focused on but in applied settings, especially when individual-level models are nested into larger ones such as markets, tractability and efficiency can become important targets.

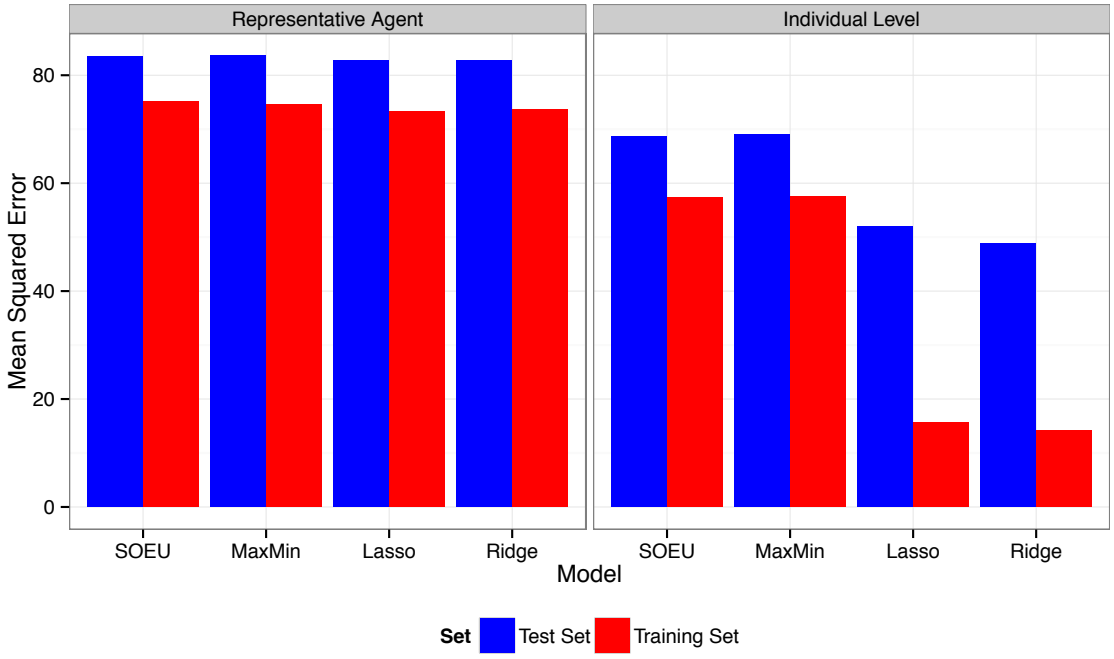


Figure 3: *ML methods outperform economic models in choice under ambiguity. This suggests that building a "plug and play" ambiguity aversion model is a fruitful direction for both theorists and experimentalists alike.*

Conclusion

We have argued that predictive power out-of-sample is an important quality for models to possess. We find that in the domain of risk, simple EU models with probability weighting do as well at predicting as machine learning algorithms. However, in the domain of ambiguity, ML outperforms economic models. This suggests that there is room for a simple "plug and play" model of ambiguity aversion that is more reflective of the "true" form of ambiguity preferences than the status quo models.

Along the way, we have introduced several techniques from machine learning including a focus on out-of-sample prediction, regularization (ie. penalizing models for complexity) and high dimensional regression. However, our standard economic models still relied on maximum likelihood estimation for fitting. An important direction for future research is to combine tools like regularization, cross validation and variable selection with more complex economic models beyond simple regressions.

Though ML methods tied with the economic models in the dimension of risk, we consider this a victory for the economic models. There is evidence that preferences about risk, time and social

decisions measured by simple economic games predict field behaviors such as insurance (Bryan 2013), cooperation (Peysakhovich et al. 2014) and taking care of one's health and other intertemporal choice behaviors (Chabris et al. 2008). Thus, models that predict well and generalize outside of the domain at hand are often more valuable than those which are simply good predictors within a particular domain. We hypothesize that a model which better captures the structural form of preferences (in our case, a better simple model of ambiguity aversion) may also predict better in field behaviors as well as work better when plugged into larger models (eg. those of markets).

We worked with EU, EUP and the two ambiguity aversion models because they are simple and have already been deployed in the literature (e.g. Levy et al. 2010, Tymula et al. 2012). We acknowledge that there are many other model choices. In the case of ambiguity, these included (but are not limited to) rank-dependent utility (Segal 1987), expected uncertain utility theory (Gul & Pesendorfer 2014), variational preferences (Maccheroni et al. 2006), smooth ambiguity models (Klibanoff et al. 2005) and others. At their core, each of these models captures ambiguity aversion by postulating that second-order uncertainty is somehow aversive. Expanding our analyses to portable parameterized versions of these models is an interesting outlet for future work. It would be especially interesting to see whether certain models are more successful on some regions of the parameter space than others, this would provide hints as to what a "final" portable ambiguity aversion model may look like.

Our results highlight the usefulness of machine learning tools for behavioral and social scientists as a benchmark for formal models as well as the importance of looking out-of-sample for evaluating model quality. In general, we argue that machine learning tools combined with the volume of data that can be gathered from the online laboratory have the potential to improve behavioral science by leaps and bounds.

Bibliography

Amir, O., Rand, DG (2012). "Economic games on the internet: The effect of $1 stakes." PLoS One **7**(2).

Bernoulli, D. (1738). “Specimen Theoriae Novae de Mensura Sortis.” Commentarii Academiae Scientarum Imperialis Petropolitane pp. 175-192

Bryan, G. (2013). “Ambiguity Aversion Decreases Demand for Partial Insurance: Evidence from African Farmers.” *Working Paper.*

Camerer, C. & Weber, M. (1992). "Recent developments in modeling preferences: uncertainty and ambiguity." Journal of Risk and Uncertainty **5**(4): 325-370.

Camerer, C. (1995). Individual decision-making. The Handbook of Experimental Economics. J. a. R. Kagel, Alvin. Princeton, Princeton University Press. **1:** 587-683.

Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Chabris, C., Laibson, D., Morris, C., Schuldt, J., Taubinsky, D. (2008) "Individual laboratory-measured discount rates predict field behavior." Journal of Risk and Uncertainty, 37(2):237-269.

Ellsberg, D. (1961). "Risk, Ambiguity and the Savage Axioms." Quarterly Journal of Economics **75**(3): 585-603.

Erev, Ido; Ert, Eyal; Roth, Alvin E. (2010). "A Choice Prediction Competition for Market Entry Games: An Introduction." Games 1, no. 2: 117-136.

Friedman, J., Hastie, T., Tibshirani, R., (2009). “The elements of statistical learning”, second ed., Springer.

Fudenberg, D. & Peysakhovich, A. (2014) "Recency, records and recaps: non-equilibrium behavior in a simple decision problem" Proceedings of the ACM 15th Annual Conference on Economics and Computation

Gilboa, I. Schmeidler , D (1989). "Maxmin expected utility with non-unique prior." Journal of Mathematical Economics **18**(2): 141-153.

Grant, S., Polak, B. & Strzalecki, T. (2009) “Second-order expected utility.” *Working paper*

Grimmer, J. (2015) “We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together.” PS: Political Science & Politics, 48(1):80-83.

Gul, F., & Pesendorfer, W. (2014). “Expected uncertain utility theory.” Econometrica, 82(1), 1-39.

Halevy, Y. (2007). "Ellsberg revisited: An experimental study." Econometrica **75**(2): 503-536.

Horton, J., Rand, DG & Zeckhauser, RJ (2011). "The online laboratory: Conducting experiments in a real labor market" Experimental Economics **14**(3): 399-425.

Kahneman, D. & Tversky, A (1979). "Prospect theory: an analysis of choice under risk." Econometrica **47**(2): 263-291.

Kahneman, D. & Tversky, A Eds. (2000). Choices, Values and Frames. Cambridge, UK, Cambridge University Press.

Klibanoff, P., Marinacci, M., & Mukerji, S. (2005). A smooth model of decision making under ambiguity. *Econometrica*, *73*(6), 1849-1892.

Knight, F (1921). "Risk, ambiguity, and profit." Boston, MA: HoughtonMifflin.

Kreps, D. (1988). Notes on the Theory of Choice. Boulder, Westview Press.

Levy, I., Snell, J, Nelson, A, Rustichini, A & Glimcher, P (2010). "Neural representation of subjective value under risk and ambiguity." Journal of Neurophysiology **103**(2): 1036-1047.

Maccheroni, F., Marinacci, M & Rustichini, A (2006). "Ambiguity aversion, robustness and the variational representation of preferences." Econometrica **74**(6): 1447-1498.

Moritz B., Zimmerman T. (2014) “Deep conditional portfolio sorts: the relation between past and future stock returns.” *Working paper.*

Naecker, J. (2014) “Using non-choice reactions to predict donation choices.” *Working paper.*

Paolacci , G., Chandler, J. (2014) “Inside the Turk: Understanding Mechanical Turk as a Participant Pool”, Current Directions in Psychological Science , 23(3):184-188

Peysakhovich, A. & Karmarkar, U. (2014) “Asymmetric effects of favorable and unfavorable information on decisions under ambiguity” *mimeo.*

Peysakhovich, A., & D. Rand. (2014). "Habits of virtue: creating norms of cooperation and defection in the laboratory." *Available on SSRN*

Peysakhovich, A., Nowak, M., Rand, D. (2014) “Humans Display a 'Cooperative Phenotype' that is Domain General and Temporally Stable” Nature Communications, Forthcoming.

Prelec, D. (1998): "The probability weighting function." Econometrica 497-527.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). “Social heuristics shape intuitive cooperation.” Nature communications, *5*.

Savage, L. (1972). Foundation of Statistics, Courier Dover Publications.

Segal, U. (1987). “The Ellsberg Paradox and Risk Aversion: An Anticipated Utility Approach.” International Economic Review 28:175–202.

Trautmann, Stefan T., and Gijs Van De Kuilen. (2013). "Ambiguity Attitudes." *Handbook of Judgment and Decision Making*.

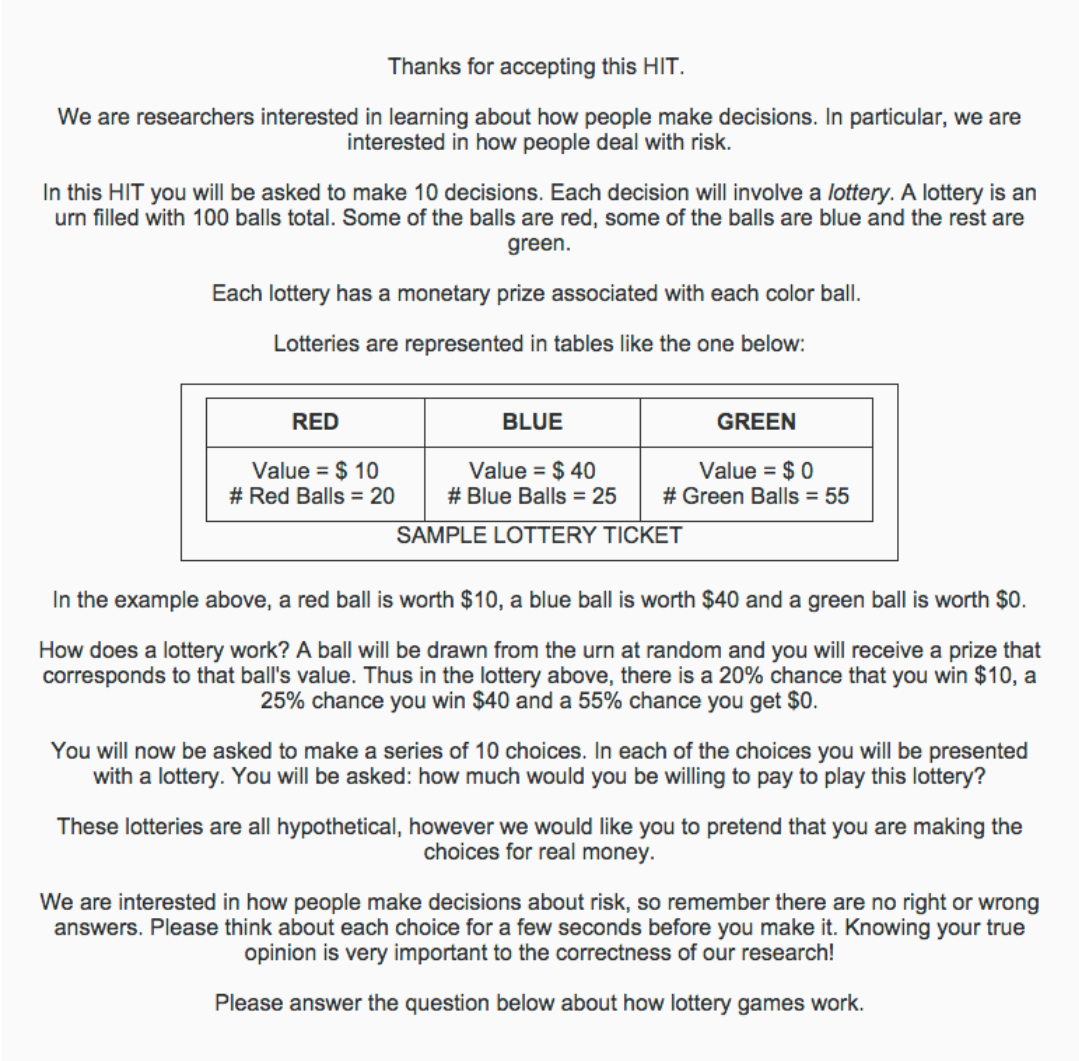
Tversky A, Kahneman D. (1992) “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” Journal of Risk and Uncertainty. 5(4):297-323.

Tymula, A., Rosenberg, L, Roy, A, Ruderman, L, Manson, K, Glimcher, P & Levy, I (2012). "Adolescents' risk-taking behavior is driven by tolerance to ambiguity." Proceedings of the National Academy of Sciences **109**(42): 17135-17140.

Varian, H. R. (2014). Big data: New tricks for econometrics. The Journal of Economic Perspectives, *28*(2), 3-27.

ONLINE APPENDIX

Appendix A: Screenshots of experiment



Thanks for accepting this HIT.

We are researchers interested in learning about how people make decisions. In particular, we are interested in how people deal with risk.

In this HIT you will be asked to make 10 decisions. Each decision will involve a *lottery*. A lottery is an urn filled with 100 balls total. Some of the balls are red, some of the balls are blue and the rest are green.

Each lottery has a monetary prize associated with each color ball.

Lotteries are represented in tables like the one below:

RED	BLUE	GREEN
Value = $ 10 # Red Balls = 20	Value = $ 40 # Blue Balls = 25	Value = $ 0 # Green Balls = 55

SAMPLE LOTTERY TICKET

In the example above, a red ball is worth $10, a blue ball is worth $40 and a green ball is worth $0.

How does a lottery work? A ball will be drawn from the urn at random and you will receive a prize that corresponds to that ball's value. Thus in the lottery above, there is a 20% chance that you win $10, a 25% chance you win $40 and a 55% chance you get $0.

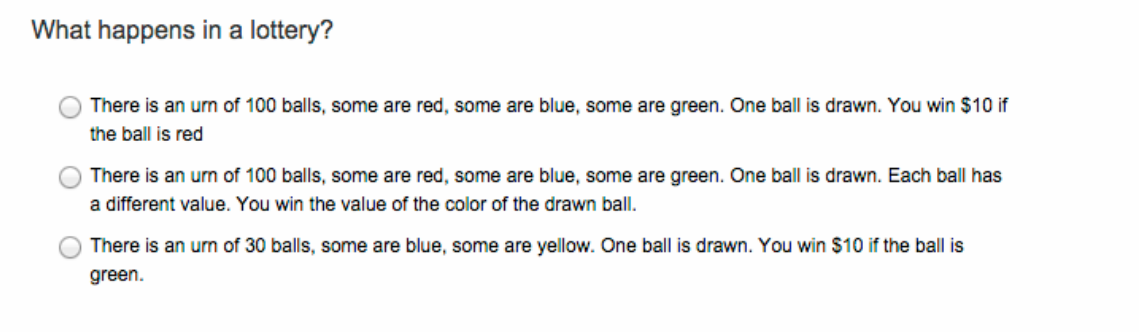
You will now be asked to make a series of 10 choices. In each of the choices you will be presented with a lottery. You will be asked: how much would you be willing to pay to play this lottery?

These lotteries are all hypothetical, however we would like you to pretend that you are making the choices for real money.

We are interested in how people make decisions about risk, so remember there are no right or wrong answers. Please think about each choice for a few seconds before you make it. Knowing your true opinion is very important to the correctness of our research!

Please answer the question below about how lottery games work.

Figure A.1 Instructions for subjects in risk version of experiment.



What happens in a lottery?

- ○ There is an urn of 100 balls, some are red, some are blue, some are green. One ball is drawn. You win $10 if the ball is red
- ○ There is an urn of 100 balls, some are red, some are blue, some are green. One ball is drawn. Each ball has a different value. You win the value of the color of the drawn ball.
- ○ There is an urn of 30 balls, some are blue, some are yellow. One ball is drawn. You win $10 if the ball is green.

Figure A.2 Understanding quiz for subjects in risk version of experiment.

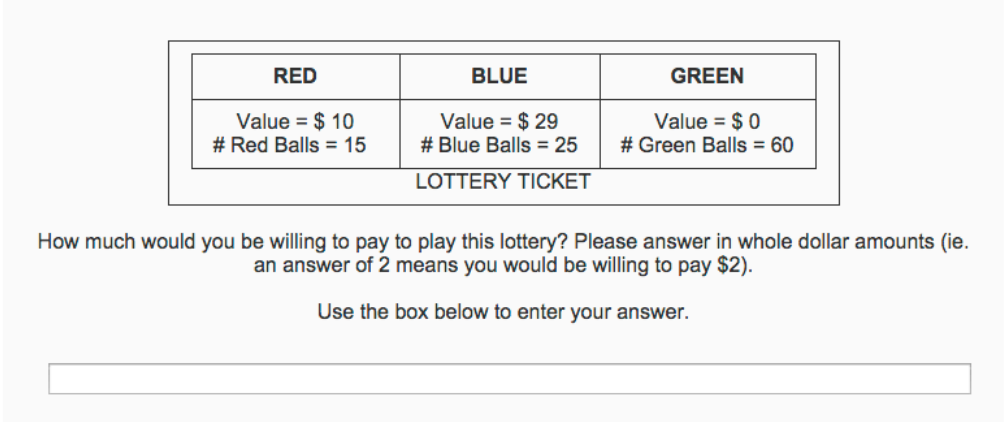


Figure A.3 Representative decision screen for subjects in risk version of experiment.

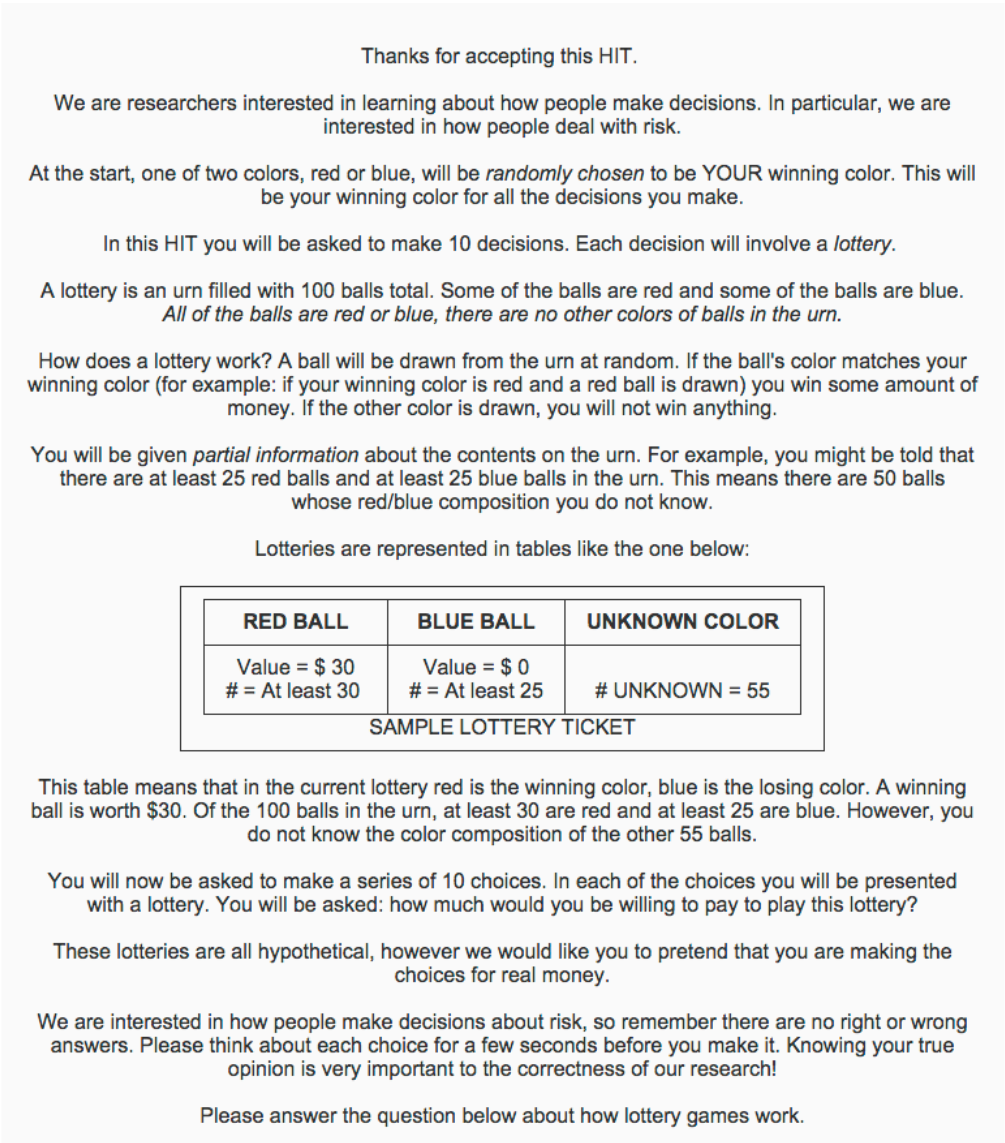


Figure A.4 Instructions for subjects in ambiguity version of experiment.

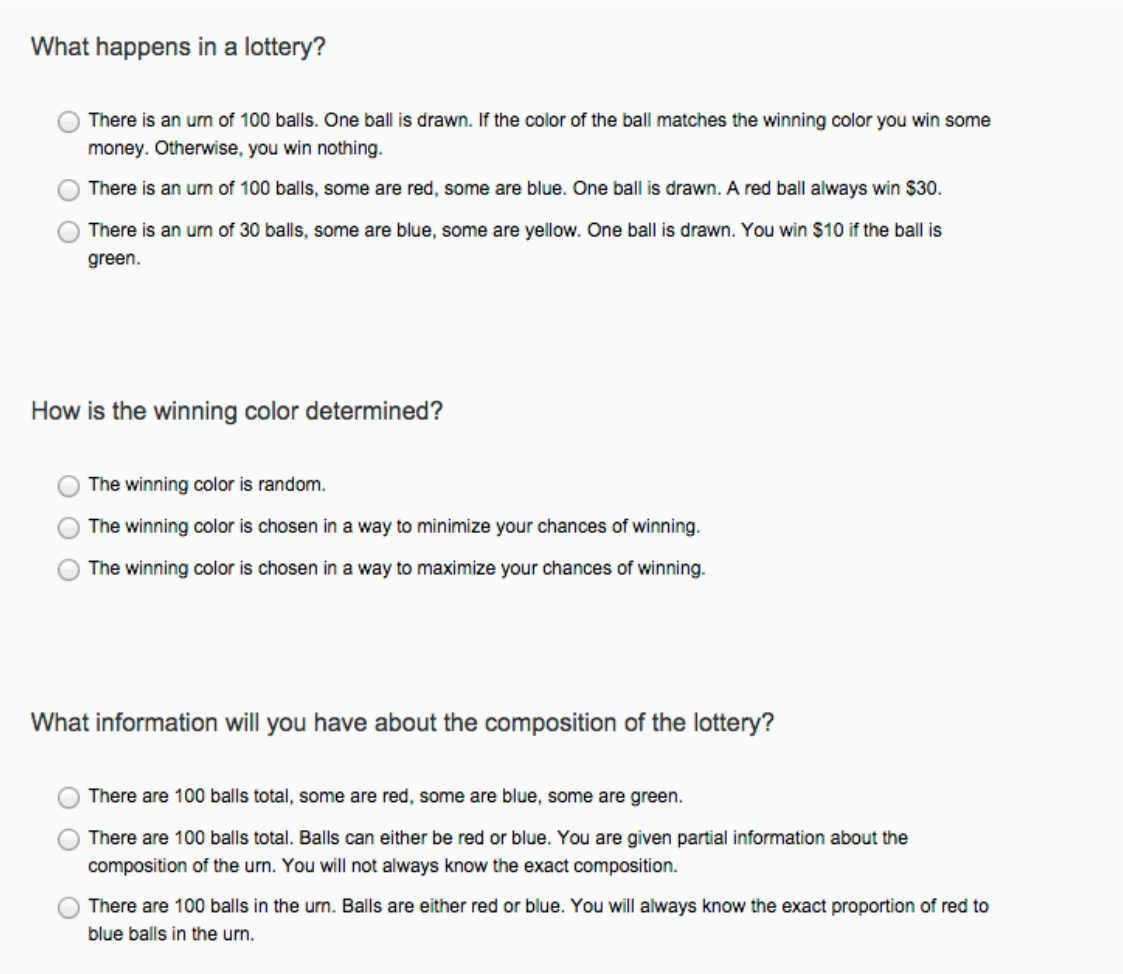


Figure A.5 Understanding quiz for subjects in ambiguity version of experiment.

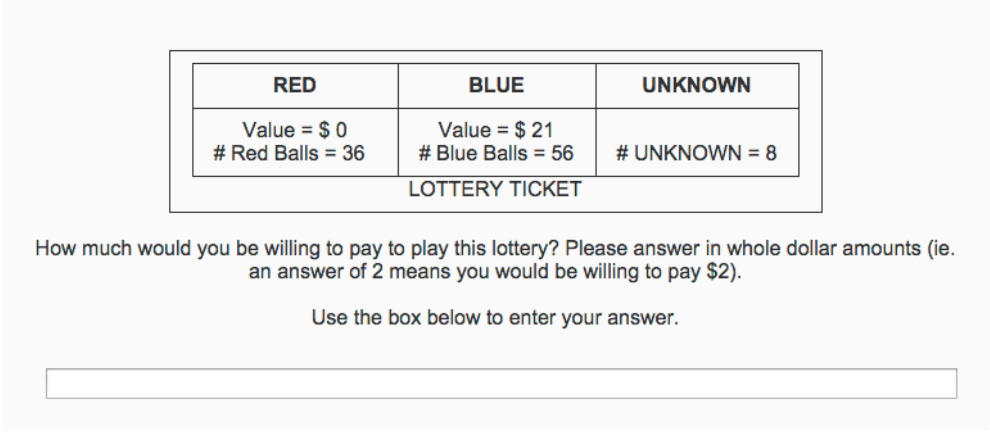


Figure A.6 Representative decision screen for subjects in ambiguity version of experiment.