ECONOMIC RECORD

THE ECONOMIC SOCIETY OF AUSTRALIA

ECONOMIC RECORD, VOL. 96, NO. 313, JUNE, 2020, 153–176

Prediction and Model Selection in Experiments*

ZACHARY BREIG

School of Economics, University of Queensland, St Lucia, Qld, Australia

This paper compares the predictive performance of several models of risk aversion and time preferences in experimental settings. Models are evaluated on the basis of out-of-sample prediction rather than in-sample fit. For preferences over risk, with the exception of very small sample sizes, allowing the estimation procedure to select between constant relative risk aversion and constant absolute risk aversion improves prediction beyond that of a single model. Moreover, adding a behavioural parameter such as disappointment aversion improves prediction further. This contrasts with time preferences, where adding the present-bias parameter worsens prediction for all sample sizes.

I *Introduction*

A large body of work in experimental and behavioural economics studies 'behavioural anomalies' and the theories that are supposed to explain them, and uses experimental and observational data to evaluate these theories. To a large extent, research in this field has focused on *testing* models, both in the theoretical and statistical sense. However, such tests do not always select a single model and are not well designed to select between models when all models are wrong.

This paper uses data from several experiments to argue for an alternative approach to choosing between models. Models are chosen on the basis of the accuracy of their predictions outside the estimation sample. The analysis focuses on deterministic models of choice, which for a given set of parameters have point predictions in these experiments. Parameters are estimated on one portion of the sample and used to predict the remainder of the data. This procedure is carried out for a number of random splits of the data, and for various sizes of the split (e.g. parameters are estimated on half the sample and used to predict the other half, then parameters are estimated on three-quarters of the sample, and used to predict the other quarter).

This methodology is applied to one experiment studying risk preferences (Choi, Fisman, Gale, & Kariv, 2007), and two studying time preferences (Andreoni & Sprenger, 2012; Augenblick, Niederle, & Sprenger, 2015). The convex budgets used in these experiments' designs allow for a clear measurement of the 'distance' between a model's predictions and a given choice.

The results compare the predictions of the constant relative risk aversion (CRRA) and the constant absolute risk aversion (CARA) models, with and without extra 'disappointment aversion' parameters. When comparing the models without these disappointment aversion parameters, the estimated CARA parameter tends to predict the subject's next choice better than CRRA for most sample sizes. On the other hand, when we add more flexibility to the model with an extra

*I thank Jim Andreoni, Joel Sobel, participants at the University of California, San Diego Theory Behavioral and Experimental lunch, and two anonymous referees for comments. I also thank those who have made their data publicly available. An earlier version of this paper was used as a chapter of my doctoral dissertation.

JEL classifications: C18, C52, D12, D81

Correspondence: Zachary Breig, The School of Economics, Level 6 Colin Clark Building (#39), The University of Queensland, St Lucia, QLD 4072, Australia. Email: z.breig@uq.edu.au

doi: 10.1111/1475-4932.12533

parameter for disappointment aversion, CRRA tends to dominate.

Estimating a disappointment aversion parameter improves prediction in the CRRA model even for small sample sizes. This may be surprising. A more complicated model necessarily improves in-sample fit for all sample sizes, but can often make prediction worse. The fact that adding parameters improves fit here suggests that researchers can obtain meaningful estimates of these parameters even with only a few observations per person, and that it might be valuable to study even more complex models in these settings. This latter conclusion is confirmed here: prediction is improved when we estimate not only the CRRA or CARA parameter for an individual, but also *which model they fit into best*. In fact, for large sample sizes the model with the highest predictive power is one in which the estimation procedure classifies a subject into which model they fall under, estimates their curvature, *and* estimates disappointment aversion parameters.

The same estimation and prediction procedure is applied to assess models of time preferences. A key issue in the literature to this point is the extent to which individuals are present biased, and how well the popular β–δ (Laibson, 1997) captures these factors. Andreoni and Sprenger (2012) found little evidence for present bias over monetary decisions; Augenblick *et al.* (2015) confirmed this result for monetary decisions, but found that present bias was observed when subjects made choices involving real effort.

The results show that estimating the present-bias parameter makes predictions worse in both experiments when considering all of the data. However, for the data from Augenblick *et al.* (2015), this masks heterogeneity in predictive power between the two types of decision problems. Estimating the β makes predictions clearly worse in monetary decisions, but when estimating on nearly the full sample, models with and without the present-bias parameter have nearly equal predictive power. The results also illustrate why researchers have estimated a wide variety of discount rates when subjects choose over time-dated money (Frederick, Loewenstein, & O'Donoghue, 2002): the prediction error for real effort decisions is much lower than the prediction error for monetary decisions.

The comparison of predictive powers of models also suggests another potential source of dynamic inconsistency: different discount rates for different sources of utility. Estimating separate discount factors for money and real effort provides improved predictions over the pooled estimate. When an agent has additively separable utility functions with differing discount rates for different goods, the agent is dynamically inconsistent. Thus, this potential source of dynamic inconsistency deserves more interest.

The methods and results presented here should be interpreted as a parallel and complementary approach to traditional methods. They give insights into the predictive power of models and demonstrate how the models interact with different amounts of data. The empirical measures that are generated are transparent and immediately interpretable. They can be used by applied modellers who are interested in choosing a model that best captures behaviour in a particular situation, and by experimenters determining the necessary quantity and source of data to estimate a particular set of preference parameters.

Section II of this paper reviews some of the related literature regarding model selection in economics, evaluating models based on prediction, and testing economic theories. Section III describes the cross-validation procedures I use, and Section IV shows the results when these procedures are applied to experiments studying risk and time preferences. Section V shows that the results are robust to a number of possible objections. Section VI concludes.

II Related Literature

Using experimental data to help distinguish between models is not a new idea. Harless and Camerer (1994) compare data from numerous studies to select between models of risk preferences. They study individual decisions in discrete choice problems from 23 different data sets. Although all the models that they study are rejected, they provide guidance about how these models trade off between parsimony and fit, favouring prospect theory, expected utility, or 'mixed fanning' (in which indifference curves fan out for unfavourable lotteries and fan in for favourable ones). Hey and Orme (1994) compare 11 different models using both standard statistical tests as well as Akaike's information criterion, and find that expected utility theory performs well, although several other models fit better (with the caveat that the economic significance of the differences is not large).

Camerer and Ho (1994) find that simple models of disappointment aversion and probability weighting fit their data much better than expected utility.

This paper is also not the first to compare models using out-of-sample prediction. Ericson, White, Laibson, and Cohen (2015) recruit subjects from Amazon.com's Mechanical Turk to make decisions that differ in framing and timing to better understand time preferences. They cross-validate the models with 100 repetitions of estimating parameters on 75 per cent of the data and predicting the remaining 25 per cent, and find that their 'intertemporal choice heuristic' performs much better than standard discounting and β–δ preferences, which perform similarly to each other. Peysakhovich and Naecker (2017) also recruit subjects from Mechanical Turk and elicit subjects' willingness to pay for various risky and ambiguous gambles. They compare the out-of-sample prediction of several economic models with data-driven machine learning models that are optimised to give better out-of-sample prediction. They find that, for all models, the representative agent assumption is a poor one (even with many fewer data points per estimated parameter, individualised parameter estimates outperform the pooled parameter estimate), and show that for risky gambles, expected utility with probability weighting performs as well out of sample as lasso and ridge regressions. On the other hand, machine learning methods outperform common economic models of ambiguous choice, suggesting that researchers have room to develop better models in this domain. While not focused directly on comparing models, Halevy, Persitz, and Zrill (2018) use a novel experimental design which estimates preferences and generates choice problems dynamically. After estimating preferences with both nonlinear least squares and what they call a 'money metric index', they compare the predictive power of the parameters estimated using these methods, and find that the latter predicts new decisions better.

This paper's motivation and results are also closely related to work by Stahl (2018), which uses the data from Hey and Orme (1994) to compare the predictive performance of several models of choice under uncertainty. The results from Stahl (2018), indicate that expected utility predicts new choices better than rank-dependent models for estimation samples that are smaller than 200, a cut-off which is much higher than is suggested by the results in this paper. These differences may be due to the nature of the decision problem faced by subjects.[1]

(i) Testing and Prediction

In most experimental work, the primary method of evaluating models is based on *testing* the models. In these cases, the experimenter finds the testable implications of a given theory, suitably adjusted for 'noise', which might be sampling variation or decision error, and carries out a statistical test in the style of Popper (1959).

This model of economic research has been embraced within the microeconomic theory literature as being the proper and scientific way of producing economic research, and decision theorists often focus on the testable implications of their models. However, there seems to be a tension in the methodological discussions about testing models. For instance, Dekel and Lipman (2010) note that a key goal of decision theory is to predict individual choice, and that a model being refuted does not imply that the model should be rejected. On the other hand, they seem to evaluate predictions primarily on the basis of whether or not the model's predictions are refuted, not how close the data are to the model's predictions. Similarly, Gilboa (2009) states that 'it is important to know that the theories have some empirical content and that given a particular mapping from theoretical objects to actual ones, a theory is not vacuous', but also that 'the question is, therefore, not whether they are right or wrong, but whether they are wrong in a way that invalidates the conclusions drawn from them'. The focus on sharp tests that arise from a model's empirical content seems to contrast with the belief that a model with strong predictive power is useful even if it is not true.

Indeed, there are numerous theories and models which have been falsified with experimental data but are still in common use today (e.g. expected utility or Nash equilibrium). Practitioners still find these concepts useful, despite the fact that they have been rejected by the data. Presumably, this is because these models are both tractable and 'good enough' to capture the phenomenon that

[1] In Choi *et al.* (2007), subjects faced convex budgets with outcomes that occurred with fixed probabilities of either one-half or one-third and two-thirds. In contrast, in Hey and Orme (1994) subjects faced binary choices in which probabilities varied within subject between one-eighth and seven-eighths.

the researcher is interested in. This argument is put forth by Simon (2007) in his 'principle of continuity of approximation', which states that 'if the conditions of the real world approximate sufficiently well the assumptions of an ideal type, the derivations from these assumptions will be approximately correct'. This article seeks to quantify and compare how closely various models approximate subjects' behaviour in experiments.

Economists have debated the necessity of testing models' assumptions for decades. In a well-known essay, Friedman (1953) discusses the ways in which economic theories are compared with data, and famously claims that the realism of a theory's assumptions cannot be used to test that theory. The essay also stresses at length that a model should be evaluated by its accuracy and usefulness *relative* to other models. Hands (1993) delves further into these topics, suggesting that Popperian methodology fails to provide rules to determine which model is better if both models have been falsified.

In addition to the theoretical debate about testing, the replication crisis has spurred additional debate around how statistical tests are carried out in practice. A large group of authors have recently come out in favour of shifting the usual significance threshold from $p = 0.05$ to $p = 0.005$ to partially overcome the issue of non-replication (Benjamin *et al.*, 2017). In response to this proposal, Lakens *et al.* (2018) have suggested that authors transparently justify the analysis choices they make, while McShane, Gal, Gelman, Robert, and Tackett (2017) suggest that both authors and journals focus on 'the totality of their data and relevant results' rather than statistical significance *per se*. The techniques presented here can be seen as a way to generate another dimension of results that allow the researcher to compare models.

This paper proposes and implements an alternative methodology to evaluate economic models. The tools presented here are not demonstrated to be optimal in any statistical sense, but are shown to be simple to implement, easy to interpret, and portable to a variety of settings.

III *Methodology*

The use of cross-validation in model selection has a long history, and there are numerous results showing the benefits and drawbacks of various methods (Arlot & Celisse, 2010). The goal of this paper is to develop a methodology that is both easily interpretable and widely applicable, and some efficiency will be sacrificed in order to obtain these goals.

A given set of subjects, indexed by $i \in \{1, \ldots, N\}$, made a series of T decisions in an experiment. When faced with a decision problem that has characteristics $x_{i,t}$, the subject chose $y_{i,t}$. Individual i's data set consists of $M_i = \{(x_{i,1}, y_{i,1}), (x_{i,2}, y_{i,2}), \ldots, (x_{i,T}, y_{i,T})\}$. A *model* maps a decision problem's characteristics, $x_{i,t}$, to a distribution over y.

The particular experiments that are focused on here are those in which subjects make choices from a convex budget, in which they allocate portions of their budget between two goods or bundles. Once a budget has been described with characteristics $x_{i,t}$, without loss of generality we can label one of the goods as good 1 and the other as good 2, such that $y_{i,t}$ can be interpreted as the proportion of their budget that they devote towards good 1.

The use of convex budgets (as compared to discrete choices) is important when comparing models on the basis of prediction. When subjects make interior choices from a convex budget problem, it is usually the case that a set of parameters is point identified. These parameters then can be used to make predictions in any other decision problem. In contrast, under discrete choice, parameters are only identified to be within some set.[2] If different parameters within this identified set lead to different choices in some other choice problem, making predictions about what the decision-maker will choose may require further assumptions.

A model and its parameters give a rule that maps a budget's characteristics to a choice. Thus, for each vector of parameters θ, such a model predicts that the agent will choose $f(x_{i,j}; \theta) \in [0, 1]$, the budget share devoted towards good 1. We can use a subject's choices and the model's predictions to get estimates for an individual's parameters, θ. In many experiments in which the researchers' goal is to estimate the parameters of a model which best describe a subject's choices, they use nonlinear least squares (NLLS). To be consistent with this literature, I do the same, which is to say that when estimating a

[2] While estimation procedures using maximum likelihood and an appropriately specified error structure give point estimates in these cases, the procedure implicitly restricts the potential set of estimated parameters to be finite.

vector of parameters $\theta_{i,j,k}$, they are chosen to solve.

$$\min_{\theta} \sum_{(x,y)\in M_i^{j,k}} (f(x;\theta)-y)^2,$$

where $M_i^{j,k}$ is the k th set of estimation budgets of size j for subject i.

In this paper each model will be estimated on many different subsets of an individual's data set. For each i, j and k, the subset of budgets that $\theta_{i,j,k}$ is being estimated on is drawn randomly. Within a set of estimations budgets, budgets are chosen without replacement (i.e. $M_i^{j,k}$ has no duplicates), but between repetitions, sets of budgets are drawn with replacement (so it is possible that $M_i^{j,k} = M_i^{j,k'}$ for $k \neq k'$). In all of the results below, models will be estimated 200 times for each estimation sample size.

The main outcome of interest is a model's predictive capability, measured in predictive mean squared error, on the portion of the sample that the model is not estimated on. Formally, for a given estimation sample size j, this can be defined as.

$$PMSE_j = \sum_{k=1}^{200}\sum_{i=1}^{N}\sum_{(x,y)\in M_i\setminus M_i^{j,k}} \frac{1}{200(T-j)N} f((x;\hat{\theta}_i^{j,k})-y)^2$$

This will be reported for all the models and experiments considered below. Patterns of the in-sample fit of each of the models are consistent across experiments, and adding parameters improves fit in the way one would expect. Thus, fit is only reported in some cases.

The data used in Sections IV and V are from several experiments studying risk and time preferences: Choi *et al.* (2007), Andreoni and Sprenger (2012), and Augenblick *et al.* (2015). In each of these experiments, subjects make a series of choices, with each choice coming from a convex budget.

I treat these experiments unfairly. The data were not meant to be used in the way I use them below, and I ignore a number of data analysis decisions that the authors use for the sake of being able to compare results across experiments. Furthermore, the results reported here rely on the specific implementation decisions made by the original authors. More information on the differences between the data analysis here and that of the original experiment can be found in Appendix I.

IV *Results*

(i) Risk Preferences

When comparing models of risk preferences, I will focus on which curvature parameters capture subjects' risk behaviours, and whether or not common behavioural parameters are useful for prediction. In particular, when considering curvature, I will compare CRRA utility functions, usually parameterised as $u(x) = (1-\rho)^{-1}x^{1-\rho}$ and CARA utility functions, parameterised as $u(x) = 1-\exp(-\rho x)$. For each of these cases, I will estimate parameters from the overall expected utility function,

$$U(x,p) = p_1u(x_1) + p_2u(x_2).$$

In addition to curvature parameters in an expected utility setting, some experiments allow for the estimation of disappointment aversion parameters. A simple parameterised version of Gul's (1991) disappointment aversion model can be seen as.

$$U(x,p) = \min\Big\{\alpha^{\mathbb{I}(x_1 \leq x_2)}p_1u(x_1) + p_2u(x_2), p_1u(x_1) + \alpha^{\mathbb{I}(x_1 \geq x_2)}p_2u(x_2)\Big\},$$

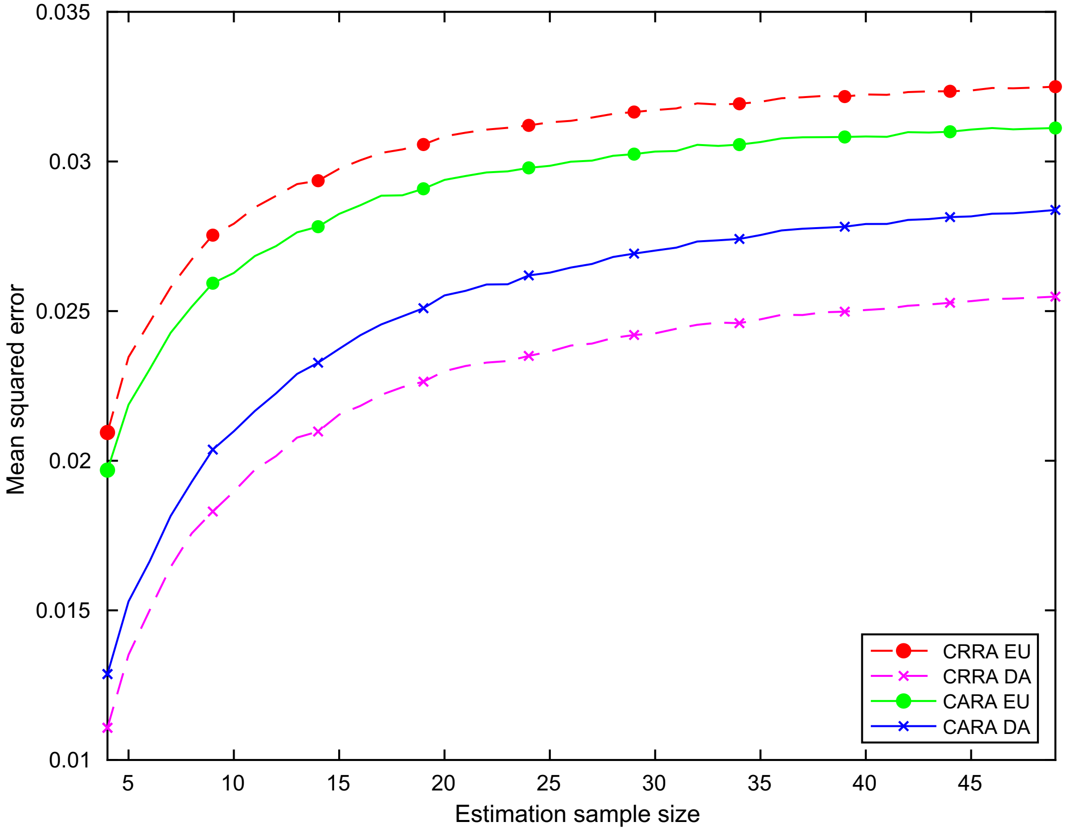
where α is the disappointment aversion parameter.[3] α is usually expected to be greater than 1, implying that the subject will place a higher weight on the utility coming from the disappointing outcome.

Results about risk preferences will use data from Choi *et al.* (2007). The experiment asks subjects to allocate their budget between two Arrow securities, exactly one of which will pay off. The subjects make the allocations under

[3] This paper estimates a disappointment aversion parameter in order for the results to be comparable to the original study by Choi *et al.* (2007). For binary lotteries, as is the case here, disappointment aversion is equivalent to a special case of probability weighting (Quiggin, 1982). Since subjects only observe lotteries with a limited set of outcome probabilities, the probability weighting function can only be identified at those values (one-half for the majority of subjects, and one-third and two-thirds for the remainder).

Figure 1

Estimation Mean Squared Error from Choi et al. *(2007). Parametric Models Fit Smaller Samples Better than they Fit Larger Samples. Allowing for Disappointment Aversion Necessarily Improves Fit for Both Models. For All Estimation Sample Sizes, Using a CARA Risk Parameter Fits the Data Better than a CRRA Risk Parameter When No Disappointment Aversion Parameter is Estimated. However, this Ranking is Reversed with the Disappointment Aversion Parameter [Colour figure can be viewed at wileyonlinelibrary.com]*

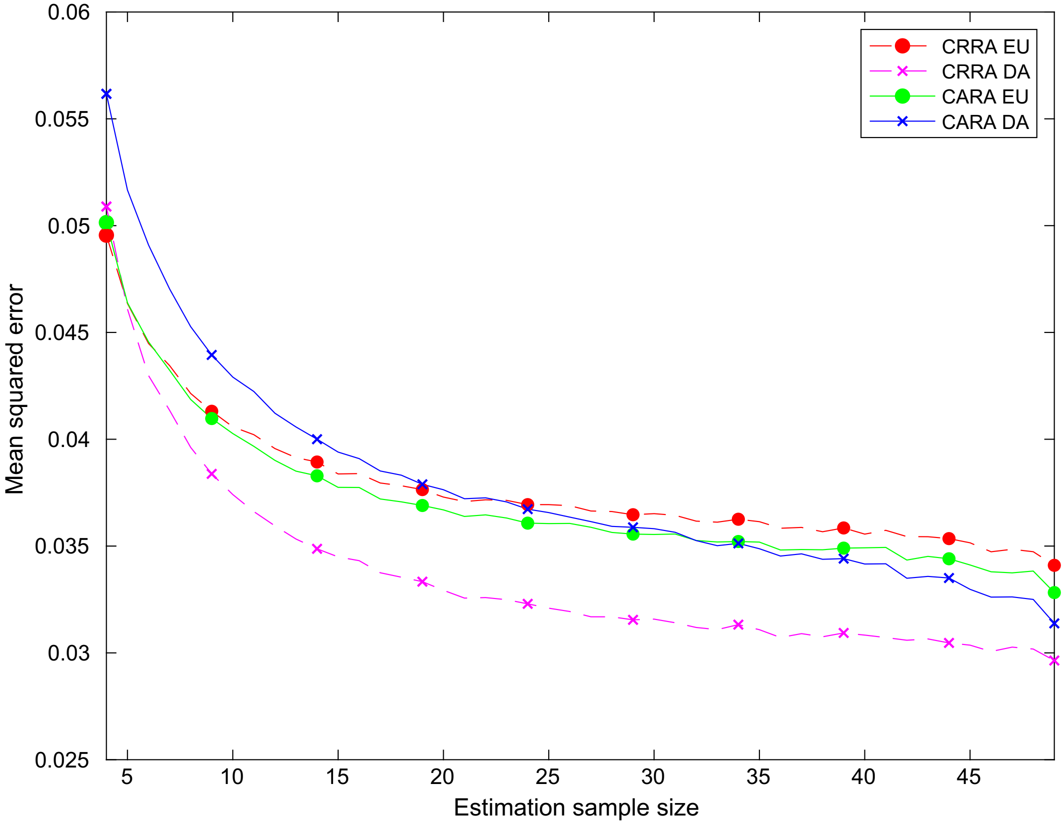


various prices. The variation of demand with prices and the total budget allows for identification of curvature parameters, and any insensitivity of relative demand to relative prices when relative prices are near 1 allows for the identification of the disappointment aversion parameter. Each subject made 50 choices, and the models' predictions will be evaluated for estimation sample sizes between 4 and 49.

The results of the estimation procedure can be seen in Figure 1. The figure shows the average estimation mean squared error for expected utility (EU) and disappointment aversion (DA) with CRRA and CARA utility functions. The x-axis refers to the size of the estimation sample, while the y-axis gives the average squared estimation error, averaged over individuals and repetitions. This graph shows that for all sample sizes, CARA

Figure 2

Prediction Mean Squared Error from Choi et al. *(2007). Out-of-sample Prediction Improves as the Size of the Estimation Sample Size Gets Larger. Estimating Disappointment Aversion Parameters can Improve Prediction Even for Relatively Small Sample Sizes [Colour figure can be viewed at wileyonlinelibrary.com]*



tends to fit the data better when no disappointment aversion parameter is estimated, but CRRA fits better with disappointment aversion. For a fixed data set on which the models are being estimated, disappointment aversion always allows the models to fit better (since the simpler model is nested in the disappointment-averse model).

The same models are compared in Figure 2, but purely based on predictions outside the estimation sample. Without disappointment aversion, CRRA gives better predictions than CARA for very small sample sizes. For larger sample sizes, CARA overtakes CRRA, confirming the inference from in-sample fit.[4] At the largest possible estimation sample sizes, CARA improves upon CRRA's prediction mean squared error by roughly 3.75% under expected utility.

Figure 2 also shows the effects of estimating disappointment aversion parameters on

[4] Differences between the in-sample fit and out-of-sample prediction indicate that estimated parameter values differ for each randomly selected sample. This is consistent with previous work studying risk elicitation tasks (Crosetto & Filippin, 2016; Loomes & Pogrebna, 2014). Gillen, Snowberg, and Yariv (2019) provide a methodology to use these noisily estimated parameters as controls in a separate regression.

prediction. Consistent with the results from Figure 1, adding disappointment aversion makes the prediction mean squared error of CRRA better than that of CARA. Furthermore, the prediction mean squared error of CRRA with disappointment aversion is lower than all of the other models even for small sample sizes, improving upon the prediction mean squared error of CRRA without disappointment aversion by roughly 13% at the largest estimation sample size. For high enough sample sizes this model *predicts* data outside the estimation sample better than the models without disappointment aversion *fit* the data they are estimated on. This is particularly interesting in light of other work which suggests that much higher sample sizes are necessary (Stahl, 2018).

Adding a disappointment aversion parameter is a simple way to extend the classic CRRA or CARA models in a way which allows for more flexibility. Other models which add flexibility such as expo-power utility (Saha, 1993) or loss aversion (Kahneman & Tversky, 1979; Kőszegi & Rabin, 2006) may also improve prediction. I discuss one such generalisation which allows the estimation process to choose between CRRA and CARA in Section V.(iii).

(ii) Time Preferences

Researchers have found renewed interest in time preferences as they have generated larger data sets which allow them to precisely estimate parameters of interest. In particular, models of dynamic inconsistency and present bias such as hyperbolic and quasi-hyperbolic discounting (Strotz, 1955; Laibson, 1997) have been the focus of significant amounts of recent research. Many papers have found evidence of present bias (Frederick *et al.*, 2002), but more recent research has questioned the robustness of this result for time-dated monetary payments (Andreoni & Sprenger, 2012), while still finding evidence for it in other kinds of choices (Augenblick *et al.*, 2015).

A simple and popular way of capturing present bias is the hyperbolic discounting model, in which the agent maximises.

$$U(c_t, c_{t+1}, c_{t+2}, \ldots) = u(c_t) + \sum_{k=1}^{\infty} \beta\delta^k u(c_{t+k})$$

where $\beta < 1$ implies present-biased decisions, while $\beta > 1$ implies future-biased decisions, and $u(\cdot)$ usually takes the form of CRRA or CARA.[5] A significant body of work has estimated these parameters using laboratory experiments in which subjects chose between earlier, smaller payments and later, larger payments.

Andreoni and Sprenger (2012) estimates these preference parameters from subjects' choices over dated monetary payments with differing interest rates.[6,7] Each subject in their experiment makes 45 choices. The range of estimation sample sizes which I will use to compare models is from 25 to 44.

The estimation mean squared error for the two models of curvature both with and without quasi-hyperbolic discounting is shown in Figure 3.[8] In this case, regardless of whether quasi-hyperbolic discounting is included, CRRA fits the data better than CARA. The quasi-hyperbolic model nests the exponential model, so fit is always improved when β is estimated.

One of the well-known results from Andreoni and Sprenger (2012) is that subjects seemed to act

[5] The classic implications of the differences between CRRA and CARA preferences relate to how consumption changes with the overall budget, holding relative prices constant. The data from Choi *et al.* (2007) are well suited to identify these differences, as they have variation in both relative prices and budgets. Andreoni and Sprenger (2012) has only minor variation in overall budget, and in Augenblick *et al.* (2015) only prices are varied. Thus, in this section differentiation between the two models relies on features of the models which may be considered secondary.

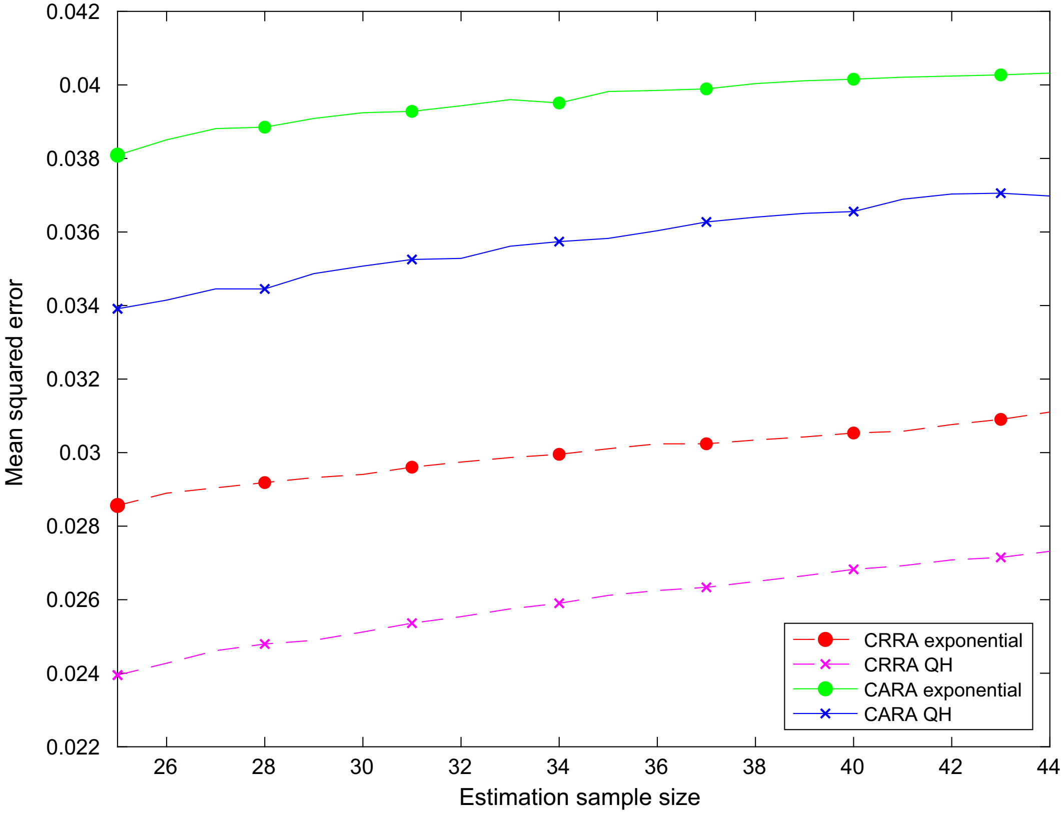
[6] It is well known that not accounting for utility function curvature may lead to a biased measure of discount rates. The convex budget set and varying interest rates used by Andreoni and Sprenger (2012) allow for a simultaneous estimation of both curvature and discounting parameters. A variety of methods have been implemented to measure curvature and overcome this potential bias (Abdellaoui, Kemel, Panin, & Vieider, 2019; Andersen, Harrison, Lau, & Rutström, 2008; Cheung, 2019; Luckman, Donkin, & Newell, 2018).

[7] The analysis here follows Andreoni and Sprenger (2012) and most of the subsequent literature in allowing for curvature parameters to vary between subjects (Abdellaoui et al., 2019; Andersen, Harrison, Lau, & Rutström, 2014; Augenblick & Rabin, 2019). One exception to this is Augenblick *et al.* (2015), which estimates discounting under the assumption that all subjects have the same cost function. Allowing for this heterogeneity gives a more realistic model, but may come at the cost of unrealistic parameter estimates.

[8] Estimation errors for all subsequent models follow a similar pattern and are omitted for brevity. They are available from the author on request.

Figure 3

Estimation Mean Squared Error for Exponential and Quasi-hyperbolic Models in Andreoni and Sprenger (2012). CRRA Curvature Parameters Always Fit the Data Better than CARA Curvature Parameters, and Adding Quasi-hyperbolic Discounting Necessarily Improves Fit [Colour figure can be viewed at wileyonlinelibrary.com]



in a way that was dynamically consistent ($\beta \approx 1$).[9] Using the same data, the prediction mean squared error in Figure 4 supports this interpretation: CRRA without quasi-hyperbolic discounting consistently outperforms CRRA with β, with a prediction mean squared error lower by almost 7 per cent at the largest estimation sample sizes. CARA with quasi-hyperbolic discounting seems to catch up to the exponential model with sample sizes larger than about 35, but never predicts appreciably better.

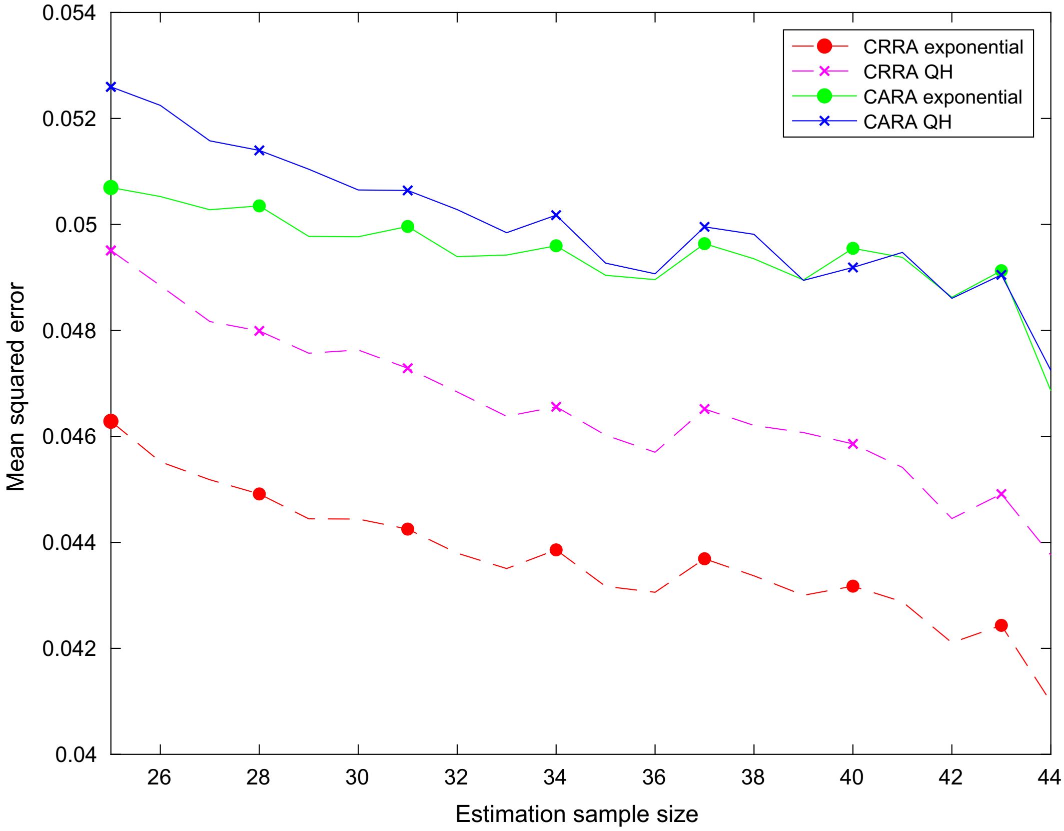
The comparison is also present when estimating the 'Stone–Geary' parameters (which can be interpreted as background consumption or consumption minima). Despite the potential of higher estimation error, another parameter improves out-of-sample fit. Even with this additional parameter, the quasi-hyperbolic model does not predict as well as the exponential.

A significant difference between the prediction error in Figure 2 and Figures 4 and 5 is the scale of the y-axis: mean squared prediction error is

[9] The finding that there is no present bias over streams of money contradicted previous experiments, but has been subsequently replicated using both convex budgets and binary choice (Andersen *et al.,* 2014; Augenblick *et al.,* 2015).

FIGURE 4

Prediction Mean Squared Error for Exponential and Quasi-hyperbolic Models in Andreoni and Sprenger (2012). Models with CRRA Curvature Parameters Predict Better than Models with CARA Parameters. Adding Quasi-hyperbolic Discounting Makes Predictions Worse [Colour figure can be viewed at wileyonlinelibrary.com]

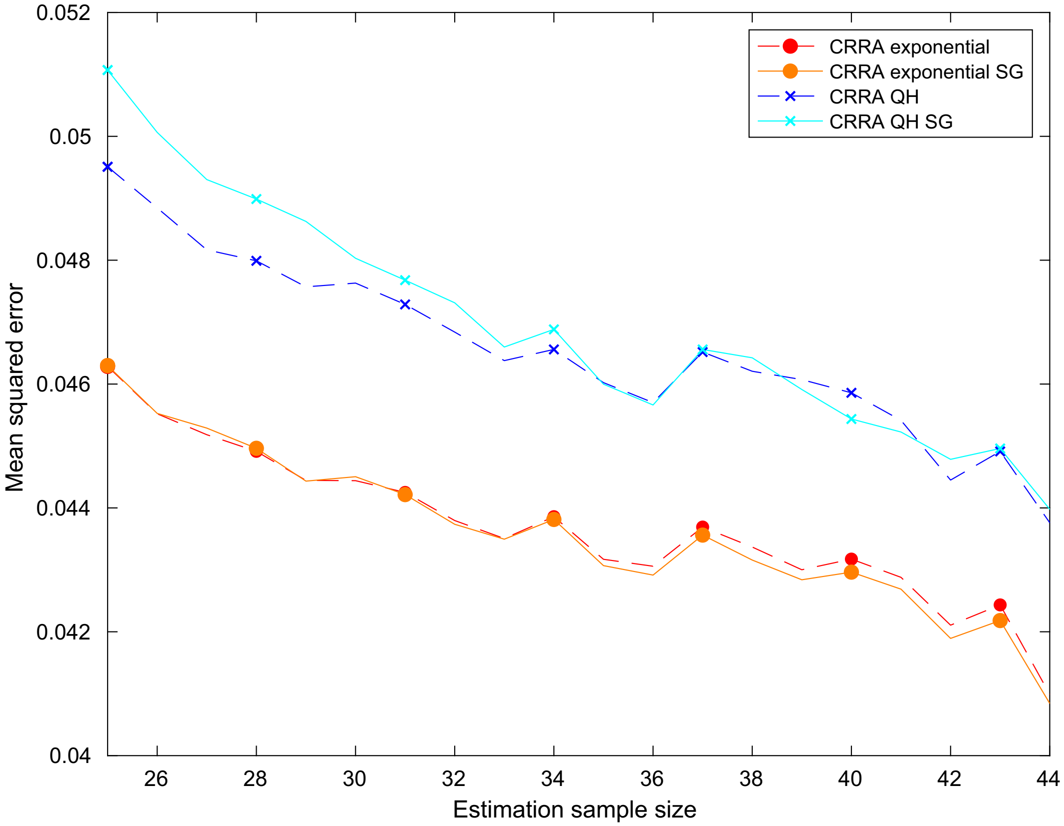


much higher when individuals make choices over dated monetary payments than when they make choices over risky prospects. There are a number of reasons why this may be true. For instance, subjects may be not be as comfortable with the type of intertemporal choice problem they are presented with, leading to more noisy decisions. Alternatively, models of risk may simply be a more accurate representation of how people choose than standard models of intertemporal preferences.

The differences in prediction error may also be due to the unique characteristics of money as the good which is being received at each date. Generally, estimated discount rates are much higher than market interest rates, implying that subjects either have credit constraints or do not consider credit markets when making their decisions (Frederick *et al.*, 2002). Furthermore, utility for time-dated monetary transfers has been estimated to be near linear (Andreoni & Sprenger, 2012; Cheung, 2019), which leads to choices on the corners of the budget constraint. An estimated model predicting the 'wrong' corner is particularly harshly punished by a convex loss function like the one used in nonlinear least squares.

Figure 5

Prediction Mean Squared Error for Models with Background Consumption in Andreoni and Sprenger (2012). Estimating Stone–Geary (SG) Background Consumption Parameters has Little Effect on Predictive Power [Colour figure can be viewed at wileyonlinelibrary.com]



In recognition of the difficulties with subjects trading off between money at sooner and later dates, more recent work has used goods that are arguably less fungible, such as time or effort. Augenblick *et al.* (2015) implement the same convex time budget design as Andreoni and Sprenger (2012), but over both monetary payments (20 decisions) and effort allocations (40 decisions). When estimating preferences over these effort allocations, Augenblick *et al.* (2015) estimate parameters from cost functions which take the form.

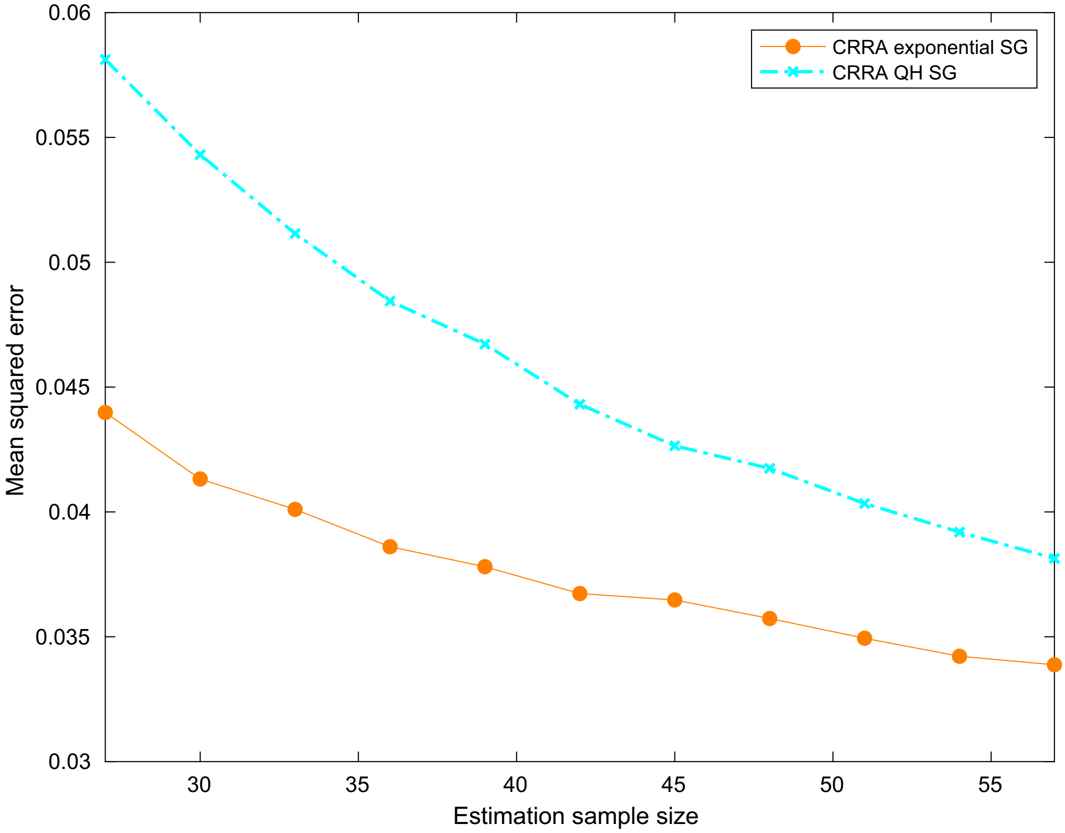
$$c(e_t, e_{t+k}) = (e_t + \omega)^\gamma + \beta^{1_{t=0}}\delta^k(e_{t+k} + \omega)^\gamma$$

where β and δ are interpreted as above, and γ, the curvature parameter on the instantaneous cost function, is expected to be greater than 1. Their results confirm Andreoni and Sprenger's finding of no time inconsistency in preferences over money, but find that in estimating quasi-hyperbolic discounting over effort provision, a significant portion of subjects have $\beta \neq 1$.

Figure 6 again compares the exponential discounting model to the quasi-hyperbolic discounting model, allowing different sets of curvature and discounting parameters for money and effort allocations. The prediction mean squared error is compared for estimation samples that include

Figure 6

Prediction Mean Squared Error from Augenblick et al. *(2015). For the Combined Data, Estimating the Present-bias Parameter β Hurts Predictive Power for All Estimation Sample Sizes [Colour figure can be viewed at wileyonlinelibrary.com]*



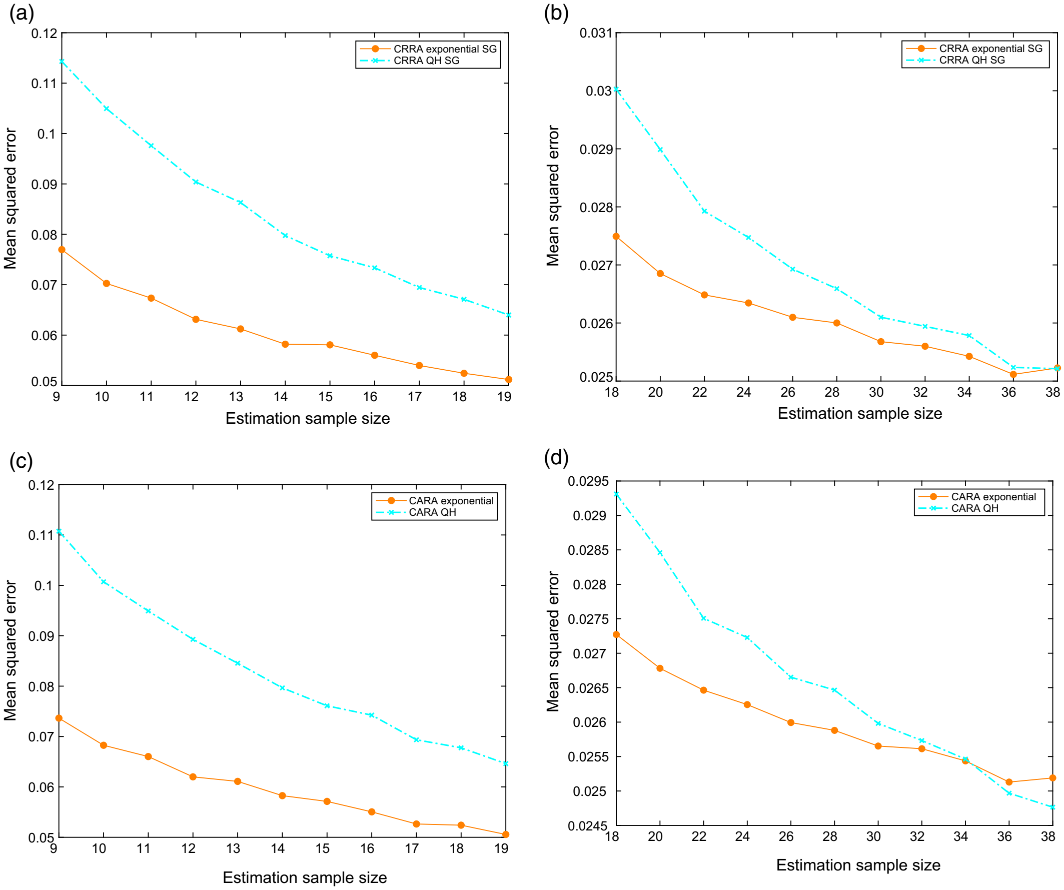
between 9 and 19 money choices for overall estimation sample sizes of between 27 and 57. The horizontal axis indicates the total number of decisions included in the estimation, and for each estimation sample size the proportion of ‘effort’ and ‘money’ decisions that were estimated on are kept the same as the overall experiment.[10]

Overall, the exponential model predicts quite well when compared to the quasi-hyperbolic discounting model, dominating it over all estimation sample sizes. However, the results from Augenblick et al. suggest that while individuals seem to act consistently over monetary allocations, they are less consistent over their effort allocations. Since Figure 6 takes a weighted average of the prediction error in these two settings, it might lead to erroneous conclusions if β–δ predicts poorly over the former but well over the latter. Thus Figure 7a,b shows the decomposition of the errors in this case. Perhaps

[10] Since the overall experiment had 20 money allocations and 40 effort allocations, this implies that estimation sample sizes of (for instance) 45 used 15 money allocations and 30 effort allocations, randomly selected.

Figure 7

Prediction MSE from Augenblick et al. *(2015) Split by Model and Decision Type. Estimating Quasi-hyperbolic Discounting Parameters Over Money Hurts Predictive Power for All Estimation Sample Sizes. Estimating the Same Parameters Over Effort Hurts Predictive Power for All but the Highest Estimation Sample Sizes [Colour figure can be viewed at wileyonlinelibrary.com]*

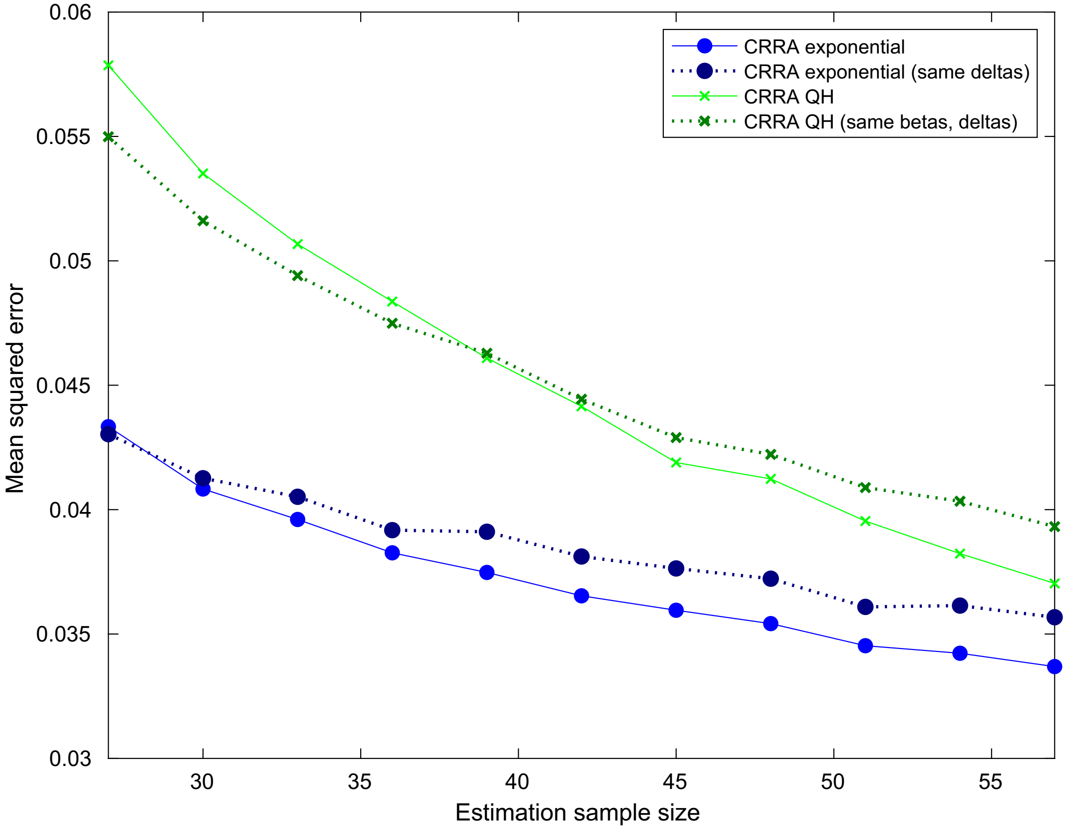


surprisingly, while the prediction mean squared error of the exponential and quasi-hyperbolic models are closer in effort allocations than in monetary allocations, the exponential model still predicts better on average for all but the largest estimation sample size, in which they have essentially equal predictive power.

In any case, the relative scales of these two panels provide an interesting comparison: these models predict time-dated effort decisions significantly better than time-dated monetary decisions. For instance, when estimating on a sample size of 18 choices over dated monetary payments, the CRRA model with exponential discounting and Stone–Geary parameters has a prediction mean squared error of 0.0524. With the same subjects, the CRRA model with exponential discounting has a prediction mean squared error of 0.0275 when estimating on a sample size of 18 choices over dated effort provision. This is consistent with a comparison across experiments: for the estimation sample sizes for which they are

FIGURE 8

Prediction MSE, for Augenblick et al. *(2015), with and without Restrictions on Equality of Discounting Parameters. Allowing Parameters to differ by the Type of Good Improves Predictive Power [Colour figure can be viewed at wileyonlinelibrary.com]*



comparable and fixing the model as CRRA with exponential discounting and Stone–Geary parameters, the prediction mean squared error from effort decisions in Augenblick et al. (Figure 7b) is less than two-thirds the prediction mean squared error from money decisions in Andreoni and Sprenger (Figure 5). In light of this and the previously discussed issues of using time-dated monetary payments, it seems that experimental methodologies using real consumption will be important in the measurement of time preferences going forward.

Figure 7c,d also estimates utility functions with an exponential loss function,

$$c(e_t, e_{t+k}) = \exp(\rho e_t) + \beta^{1_{t=0}}\delta^k \exp(\rho e_{t+k}),$$

and CARA utility, for completeness. The two types of curvature have almost equal predictive power over money decisions and effort decisions without quasi-hyperbolic discounting. On the other hand, using a CARA curvature parameter seems to improve predictions over CRRA when the β–δ model is applied to effort decisions. Indeed, of the four models, CARA with quasi-hyperbolic discounting predicts effort decisions best for high sample sizes.

A unique feature of Augenblick et al's data is that it contains information about subjects' preferences regarding two *different* valuable goods. They estimate present-bias parameters for effort and money in different regressions, getting separate estimates for the two, and then show that they are uncorrelated. A natural question is how the decision to estimate these parameters separately affects the prediction power of these models. Figure 8 shows the results of these regressions. The solid lines are the prediction mean squared error of the model in which the effort and money discounting parameters are allowed to be different, while the dotted lines restrict them to be the same. For both the exponential and the β–δ model, allowing different discounting factors *improves* prediction for almost all sample sizes, despite requiring the estimation of more parameters.

The possibility of different goods being discounted differently has been discussed in the literature, but these results suggest that the topic deserves more interest. Banerjee and Mullainathan (2010) demonstrate that this sort of discounting leads to time inconsistency (even when $\beta = 1$) and preference reversals which have elsewhere been attributed to present bias. Furthermore, the result here is in accordance with previous work showing that individuals have good-specific discount rates (Winer, 1997; Odum & Rainaud, 2003; Ubfal, 2016). Analysing new sources of data with good-specific discount rates is likely to be a fruitful line of research, although it remains to be seen how much is lost by assuming additively separable utility over time.

V *Robustness*

(i) Statistical Significance

The results shown above do not address the statistical significance of the difference between models' predictive capabilities. This was deliberate: since a primary goal is to select the set of preferences that will be used by applied modellers, one must choose the 'winner', whether or not the results are significant in the statistical sense.

With this caveat in mind, it will still be useful to those running experiments to have a sense of where more research needs to be done to provide a definitive answer. As a first measure, it is reassuring that there is substantial consistency within and across experimental data sets. If a model predicts substantially better when estimated on a sample of size 40, it generally also predicts better when estimated on sample sizes of 39 or 41. Furthermore, in the one direct comparison across experiments that can be made here, the money prediction error Figure 7a is similar to the prediction error from Figure 4 when comparing similar estimation sample sizes. Beyond these points, I also provide two measures of the confidence in these results.

The first measure is a confidence set for each model's mean squared prediction error. Since, for each estimation sample size, a different estimation sample was drawn 200 times, a natural measure of the variance of the prediction error is provided by the 5th and 95th percentiles of the prediction error of these draws. Figure 9 shows these confidence sets for exponential and quasi-hyperbolic discounting under CRRA curvature. In general, these confidence sets have significant overlap. The primary reason for this is that within a given subset of the data, the prediction error of models is highly correlated: some data sets are harder to predict than others.

To account for this correlation, I also calculate *t*-statistics from the matched-pairs *t*-test for the difference in means, which accounts for the correlation in difficulty of prediction for a given draw of the data. When this is taken into account, equality of means is rejected at very low *p*-values. For instance, when comparing the difference in mean prediction mean squared error of the expected utility models, as in Figure 9a, the only estimation sample sizes for which equality of means is *not* rejected at the 5% level are 5 and 6. Equality is always rejected when testing difference in means between the disappointment aversion models, as in Figure 1b, and the highest *p*-value is less than 0.001.

(ii) Akaike information criterion

In addition to predictive measures such as those used here, another common method of model selection is the use of information criteria, the most common of which are the Akaike information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978). Since here we are using NLLS to estimate parameters, the most natural measure is the AIC, which in the case of least squares is usually defined as.

$$AIC = n\ln(MSE) + 2k \qquad (1)$$

where n is the number of observations, MSE stands for estimation mean squared error, and k is the number of parameters. In our case, the total

FIGURE 9

Means of Prediction MSE from Choi et al *(2007) with Confidence Sets Generated from the 5th and 95th Percentiles. Even though Matched Pair t-tests Reject Equality for Most Estimation Sample Sizes, these Confidence Sets Generally Overlap [Colour figure can be viewed at wileyonlinelibrary.com]*

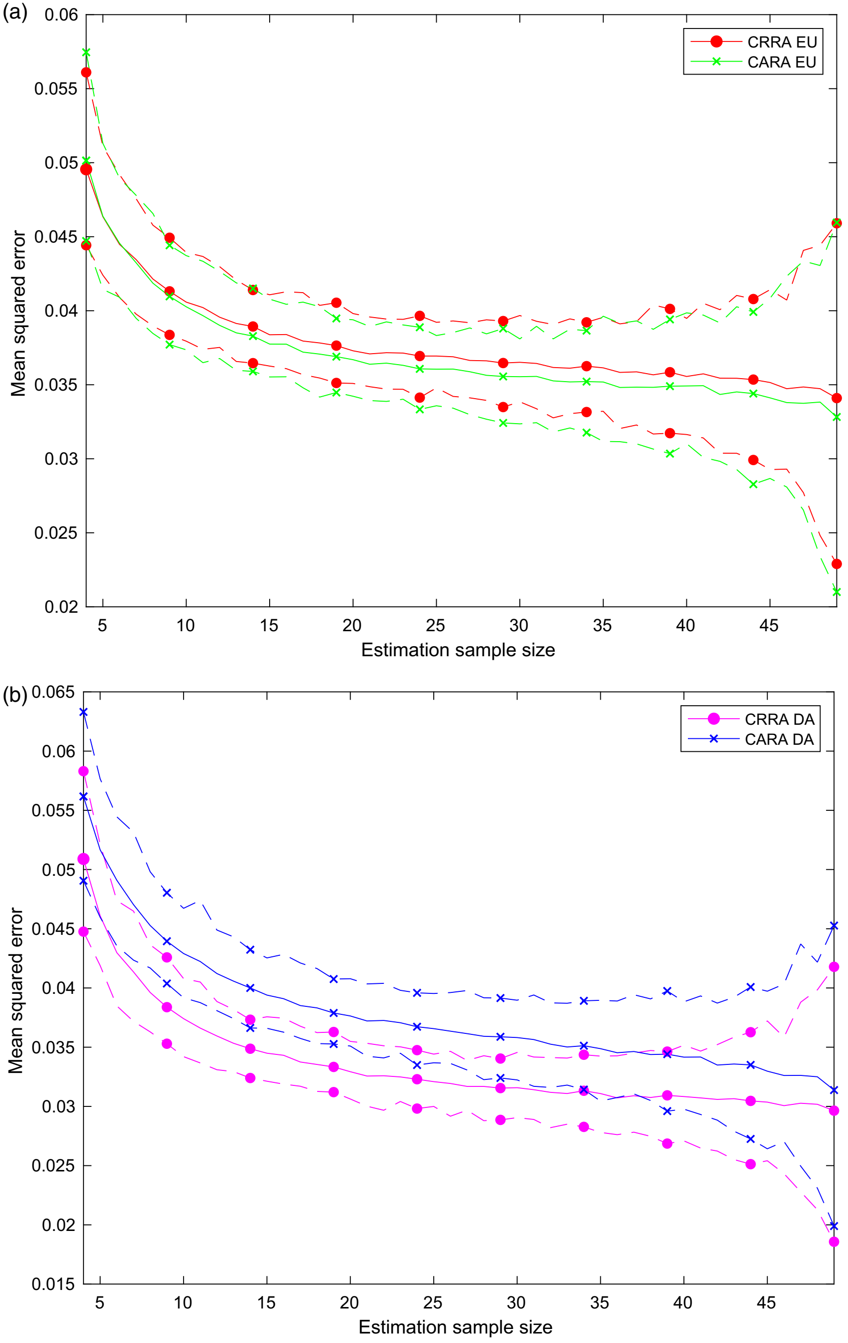


Table 1
AIC Versus Prediction Mean Squared Error from Augenblick et al. *(2015). Here, the Rankings Implied by the AIC Coincide With Those Given by Prediction Mean Squared Error When Estimating on Nearly the Full Sample*

	CRRA EU	CRRA DA	CARA EU	CARA DA
Prediction MSE rank	4	1	3	2
AIC rank	4	1	3	2

Table 2
AIC Versus Prediction Mean Squared Error from Augenblick et al. *(2015). The AIC and Prediction Mean Squared Error Give Different Rankings When Estimating on Nearly the Full Sample*

	CRRA Exp SG	CRRA QH SG	CARA Exp	CARA QH
Prediction MSE rank	2	4	1	3
AIC rank	4	2	3	1

number of observations is the number of subjects in the experiment multiplied by the number of decisions, and the number of parameters is the number of subjects in the experiment multiplied by the number of parameters in the model being studied.

Since the purpose of the AIC is to select the best model for the full sample, the most natural comparison between AIC and the methods used in this paper is to consider the rankings provided by AIC versus the rankings given by mean squared prediction error on the largest possible estimation sample size.

Tables 1 and 2 give the rankings implied by these two measures for two of the experiments considered above. These measures only sometimes coincide; in the data from Choi *et al.* (2007), prediction mean squared error and the AIC would give exactly the same ranking of models. In the data from Augenblick *et al.* (2015), they substantively differ. For prediction mean squared error, the exponential CARA model is the best and quasi-hyperbolic CRRA with Stone–Geary parameters is the worst. For AIC the ranking is quite different: the quasi-hyperbolic CARA model is the best and exponential CRRA model with Stone–Geary parameters is the worst.

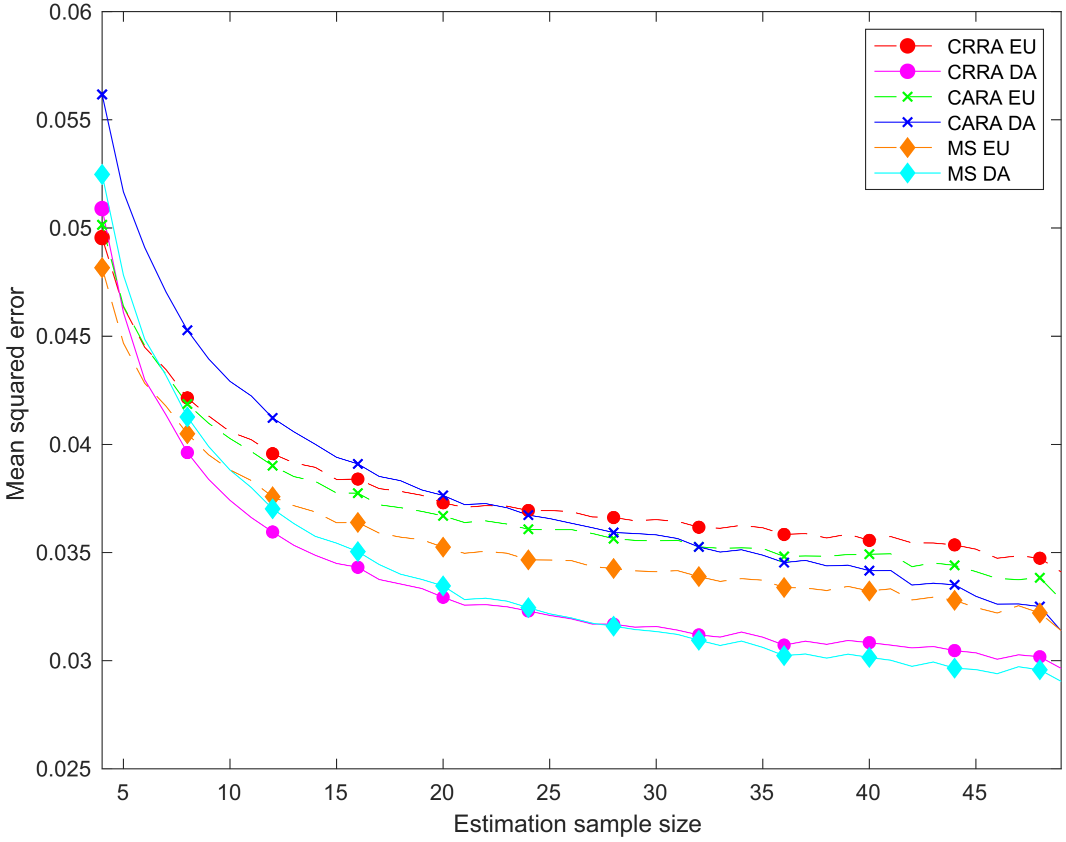
The primary benefit of using the AIC (as compared to cross-validation measures) is that it is very easy to calculate: the estimation mean squared error had to have been calculated in the estimation process anyway, and the subsequent calculation of the AIC is trivial. In comparison, cross-validation is much more demanding to compute: it requires estimating each of the models many times, which may involve substantial coding and computation time.

On the other hand, there are several qualities of cross-validation which make it superior to the AIC. The first is interpretability: the AIC does not have an interpretation itself, but is only useful when comparing models. Mean squared prediction error, on the other hand, is easily interpretable as a measure of the distance between predicted and actual decisions. Related to this point, it may not be easy to compare the results of an AIC calculation in one experiment to that from another; it is easy to compare the results of prediction error across experiments, and if the experiments use convex budgets, they need not even measure preferences in the same choice domain.

Second, the prediction error is more closely tied to the purpose of estimating these preferences. The parameters which are estimated using experimental choices are supposed to capture what subjects will do *outside* the lab, and in the past have been evaluated based on their correlation with real-life decisions (Meier & Sprenger, 2010; Fisman, Jakiela, Kariv, & Markovits, 2015; Fisman, Jakiela, & Kariv, 2017). Mean squared prediction error is in some sense an intermediate step between the goals of fitting a model to experimental choices and predicting real-world behaviour. The AIC, on the other hand, can only be interpreted this way in so far as one can argue that it reduces estimation error, and the particular

FIGURE 10

Prediction MSE, Including Model Selection, Choi et al. *(2007). Allowing the Estimation Procedure to Select Which Model a Subject Uses to Choose Improves Predictive Power for Most Sample Sizes, and Adding a Disappointment Aversion Parameter on Top of this Improves it Further for Large Enough Estimation Sample Sizes [Colour figure can be viewed at wileyonlinelibrary.com]*



formula chosen for the AIC is mainly used for historical reasons.[11]

(iii) Model Heterogeneity

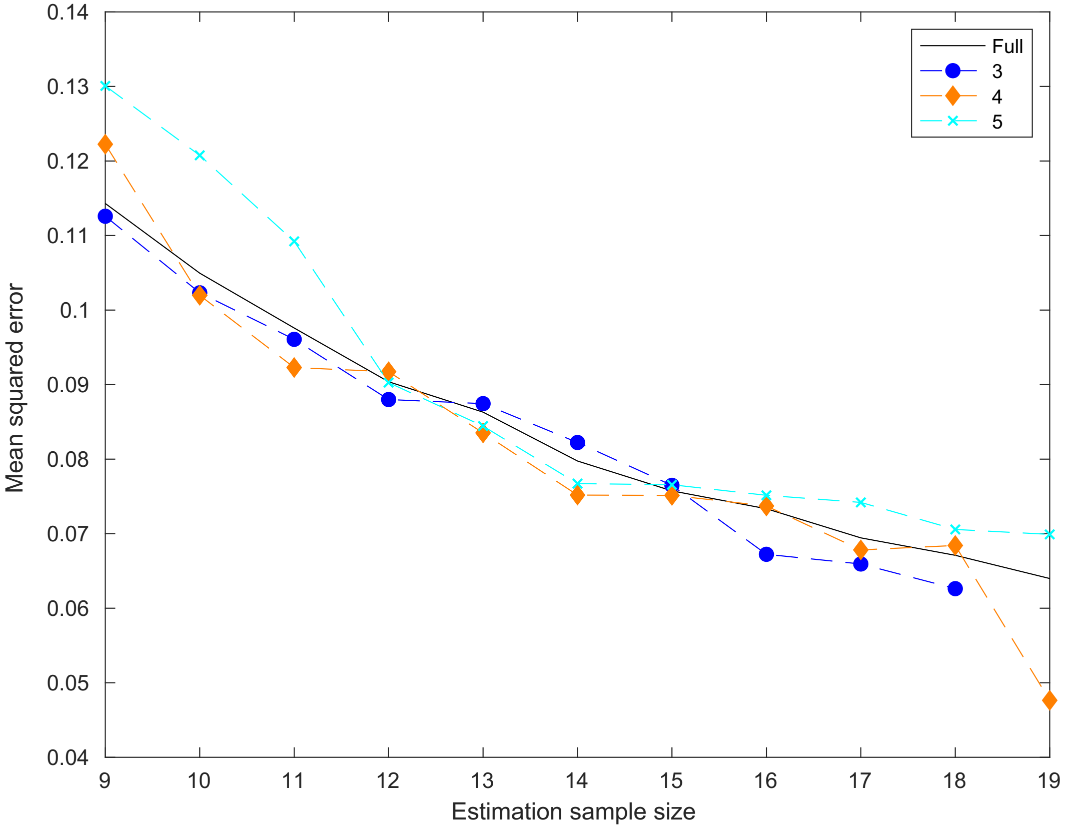
The estimation procedures used to this point have assumed that each subject's behaviour is explained by the same utility function, but with different parameters. Previous research has demonstrated that there is significant heterogeneity in preference parameters arising in a wide variety of settings. With this in mind, one might ask whether different individuals' decisions might be explained not only by different parameter values in utility functions, but also by different families of utility functions themselves.

To answer this question I treat the model itself as another parameter to be estimated. Thus, I estimate parameters of each model using the data from a subset of an individual's budgets, and

[11] One might argue that mean squared error is also an arbitrary measure of how well a model fits, and this is true. However, mean squared error is used in both cases, and to use the AIC one has to make the extra arbitrary decisions that give the formula in Equation (1).

Figure 11

Prediction MSE for Money Decisions in Augenblick et al. *(2015) for All Estimation Samples, and for Estimation Samples Containing the Given Number of "Only Future" Decisions. High Prediction MSE is not the Result of the Randomization Procedure Selecting an Unrepresentative Sample [Colour figure can be viewed at wileyonlinelibrary.com]*

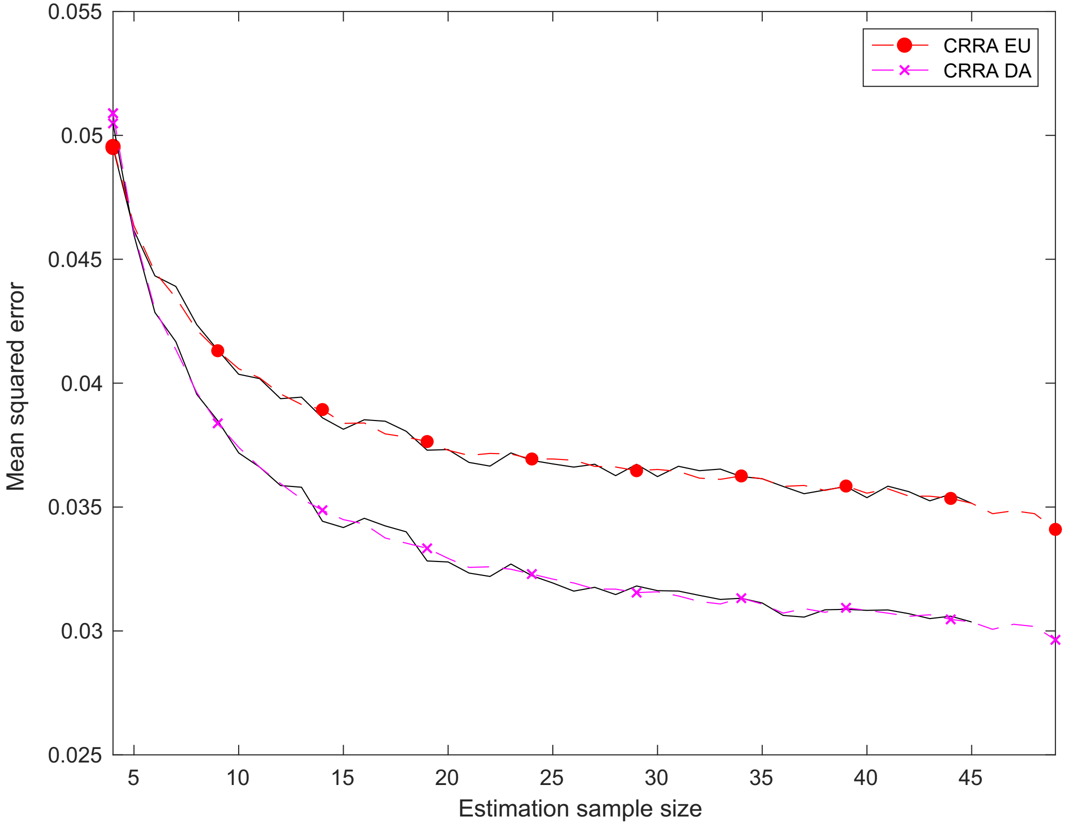


choose the model which has the lower mean squared estimation error. The prediction error is then calculated using the parameters from the chosen model from that individual.

The results of this procedure are shown in Figure 10 for Choi *et al.* (2007). Since the mean squared estimation error is necessarily lower for the models with disappointment aversion, I execute the procedure separately for expected utility and disappointment-averse models. Perhaps surprisingly, this heterogenous model procedure without disappointment aversion uniformly outperforms the other expected utility models, and the heterogeneous model procedure with disappointment aversion improves on the other disappointment-averse models with sufficient estimation sample size.[12] This suggests that allowing for heterogeneity in the family of risk preferences in addition to heterogeneity in parameters might better rationalise individuals' decisions.

[12] Another interesting result of this procedure is the classification of each subject into one of the two types: CRRA or CARA.

FIGURE 12
Prediction MSE, Including Constant Comparison in Black, Choi et al. *(2007). Results do not Change if the Prediction Sample Size is Held Constant Rather than Changing with the Estimation Sample Size [Colour figure can be viewed at wileyonlinelibrary.com]*



(iv) Identification and Experimental Design

The cross-validation procedure carried out above ignores important features of experimental design. Researchers design these experiments with a particular mix of choice problems to identify parameters of interest. For instance, in an experiment designed to estimate the present-bias parameter β, subjects face a number of choices in which they trade off between the present and the future, as well as decisions in which they trade off between two future points. When the cross-validation procedure splits a subject's decisions into the estimation and prediction samples, it may do so in a way that does not include enough decisions of a given type. If this is the case, a researcher using cross-validation might be overly pessimistic about the more complex model.

It is likely that this problem will be worst when the ratio of parameters to identifying points is high. This is because if only a few of the points that can identify a parameter are being used in the estimation procedure, this estimate is likely to be very noisy. Of the models and data sets studied here, the results regarding the prediction error in money decisions for the quasi-hyperbolic model with Stone–Geary background consumption seems to be the most likely case in which it will arise. The 'money' portion of the model has four parameters (β, δ, γ, and ω), estimated using at most 20 data points. Furthermore, a maximum of five of these data points are decisions including only 'future' monetary payment, which are critical for the separate identification of β and δ.

This issue does not seem to drive the results found above. To show that it is does not, Figure 11 disaggregates the prediction mean squared error by how many 'future' decisions are in the estimation sample, and includes the prediction mean squared error from the full sample for comparison. Each of the dashed lines is an average of the prediction mean squared error, where the number of 'only future' choices is held constant. The main result from Figure 11 is that the prediction mean squared error is *not* being driven by non-representative estimation samples, in which there are 0, 1, or 2 'only future' future decisions. Instead, the prediction mean squared error the full sample are generally close to those which estimate using 3 to 5 of these 'only future' decisions.

(v) Constant Prediction Sample

In all of the analysis above, as one increases the number of budgets used in the estimation, the number of budgets used for prediction gets smaller. When a model is estimated on sample size j and the full experiment generated T observations per subject, the prediction mean squared error is calculated by finding the mean squared prediction error on T-j observations. One might imagine that the changing size of the prediction sample could be influencing the results.

An alternative would be to keep the size of the prediction sample constant. Obviously, if the prediction sample size is held constant at (for instance) 5, the largest possible estimation sample size is T-5. The results of completing this exercise for the CRRA with and without disappointment aversion can be seen in Figure 6. Here, the mean squared prediction error on a constant sample size of 5 is overlaid on the mean squared prediction error on the complement of the estimation data. The mean squared prediction error on a constant sample size is almost exactly equal to the mean squared prediction error on the changing sample size. In fact, the former seems to be a slightly noisier version of the latter; and this is true for all of the models estimated in this paper. Thus, one can be confident that the changing prediction sample size is not driving any of the above results (Figure 12).

VI Conclusion

Economic experiments are well posed to allow economists to select between models of behaviour. This paper presents a method to select between these models on the basis of out-of-sample prediction.

The procedure is applied to data from several experiments that elicited risk and time preferences. When studying subjects' risk preferences, more complicated models that allow for disappointment aversion and broadly different risk behaviour predict better than simpler models. Furthermore, these models tend to predict better even when only estimating on a few data points. When applied to decisions over time, the methodology shows that exponential discounting predicts as well as quasi-hyperbolic discounting, and predicts better for smaller sample sizes. However, the models that predict well are also time inconsistent, since they have different discount rates for different goods.

Prediction is presented as a complementary approach to standard methods of testing. Its outcomes are easily interpretable, and it does not suffer from the possibility of rejecting all competing models. In addition to providing guidance to applied modellers, the results are also useful for the empirical and experimental researchers: they implicitly show the amount of data needed to convincingly estimate parameters from a given model.

REFERENCES

Abdellaoui, M., Kemel, E., Panin, A. and Vieider, F.M. (2019), 'Measuring time and risk preferences in an integrated framework', *Games and Economic Behavior*, **115**, 459–469.

Akaike, H. (1973), 'Information theory and an extension of the maximum likelihood principle', in Petrov, B. and Csáki, F. (eds), *Second International Symposium on Information Theory*. Akadémiai Kiadó, Budapest; 267–281.

Andersen, S., Harrison, G.W., Lau, M.I. and Rutström, E.E. (2008), 'Eliciting risk and time preferences', *Econometrica*, **76** (3), 583–618.

Andersen, S., Harrison, G.W., Lau, M.I. and Rutström, E.E. (2014), 'Discounting behavior: A reconsideration', *European Economic Review*, **71**, 15–33.

Andreoni, J. and Sprenger, C. (2012), 'Estimating time preferences from convex budgets', *American Economic Review*, **102**, 3333–3356.

Arlot, S. and Celisse, A. (2010), ‘A survey of cross-validation procedures for model selection’, *Statistics Surveys*, **4**, 40–79.

Augenblick, N. and Rabin, M. (2019), ‘An experiment on time preference and misprediction in unpleasant tasks’, *Review of Economic Studies*, **86**, 941–975.

Augenblick, N., Niederle, M. and Sprenger, C. (2015), ‘Working over time: Dynamic inconsistency in real effort tasks’, *Quarterly Journal of Economics*, **130**, 1067–1115.

Banerjee, A. and Mullainathan, S. (2010), ‘The shape of temptation: Implications for the economic lives of the poor’, NBER Working Paper No. 15973, National Bureau of Economic Research, Cambridge, MA.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C.D., Clyde, M., Cook, T.D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., Fehr, E., Fidler, F., Field, A.P., Forster, M., George, E.I., Gonzalez, R., Goodman, S., Green, E., Green, D.P., Greenwald, A.G., Hadfield, J.D., Hedges, L.V., Held, L., Hua Ho, T., Hoijtink, H., Hruschka, D.J., Imai, K., Imbens, G., Ioannidis, J.P.A., Jeon, M., Jones, J.H., Kirchler, M., Laibson, D., List, J., Little, R., Lupia, A., Machery, E., Maxwell, S.E., McCarthy, M., Moore, D.A., Morgan, S.L., Munafó, M., Nakagawa, S., Nyhan, B., Parker, T.H., Pericchi, L., Perugini, M., Rouder, J., Rousseau, J., Savalei, V., Schönbrodt, F.D., Sellke, T., Sinclair, B., Tingley, D., Van Zandt, T., Vazire, S., Watts, D.J., Winship, C., Wolpert, R.L., Xie, Y., Young, C., Zinman, J. and Johnson, V.E. (2017), ‘Redefine statistical significance’, *Nature Human Behaviour*, **2**, 6–10.

Camerer, C.F. and Ho, T.-H. (1994), ‘Violations of the betweenness axiom and nonlinearity in probability’, *Journal of Risk and Uncertainty*, **8**, 167–196.

Cheung, S.L. (2019). Eliciting utility curvature in time preference. *Experimental Economics*. Retrieved from https://doi.org/10.1007/s10683-019-09621-2

Choi, S., Fisman, R., Gale, D. and Kariv, S. (2007), ‘Consistency and heterogeneity of individual behavior under uncertainty’, *American Economic Review*, **97**, 1921–1938.

Crosetto, P. and Filippin, A. (2016), ‘A theoretical and experimental appraisal of four risk elicitation methods’, *Experimental Economics*, **19**, 613–641.

Dekel, E. and Lipman, B.L. (2010), ‘How (not) to do decision theory’, *Annual Review of Economics*, **2**, 257–282.

Ericson, K.M.M., White, J.M., Laibson, D. and Cohen, J.D. (2015), ‘Money earlier or later? Simple heuristics explain intertemporal choices better than delay discounting does’, *Psychological Science*, **26**, 826–833.

Fisman, R., Jakiela, P., Kariv, S. and Markovits, D. (2015), ‘The distributional preferences of an elite’, *Science*, **349**, aab0096–aab0096.

Fisman, R., Jakiela, P. and Kariv, S. (2017), ‘Distributional preferences and political behavior’, *Journal of Public Economics*, **155**, 1–10.

Frederick, S., Loewenstein, G. and O’Donoghue, T. (2002), ‘Time discounting and time preference: A critical review’, *Journal of Economic Literature*, **40**, 351–401.

Friedman, M. (1953), The methodology of positive economics. In *Essays in Positive Economics*. University of Chicago Press, Chicago, IL, 3-16.

Gilboa, I. (2009). *Theory of Decision Under Uncertainty*. Cambridge University Press, Cambridge, UK.

Gillen, B., Snowberg, E. and Yariv, L. (2019), ‘Experimenting with measurement error: Techniques with applications to the Caltech Cohort Study’, *Journal of Political Economy*, **127**, 1826–1863.

Gul, F. (1991), ‘A theory of disappointment aversion’, *Econometrica*, **59**, 667–686.

Halevy, Y., Persitz, D. and Zrill, L. (2018), ‘Parametric recoverability of preferences’, *Journal of Political Economy*, **126**, 1558–1593.

Hands, D.W. (1993), ‘Popper and Lakatos in economic methodology’, *Rationality, Institutions, and Economic Methodology*, **2**, 61.

Harless, D.W. and Camerer, C.F. (1994), ‘The predictive utility of generalized expected utility theories’, *Econometrica*, **62**, 1251–1289.

Hey, J.D. and Orme, C. (1994), ‘Investigating generalizations of expected utility theory using experimental data’, *Econometrica*, **62**, 1291–1326.

Kahneman, D. and Tversky, A. (1979), ‘Prospect theory: An analysis of decision under risk’, *Econometrica*, **47**, 263–291.

Kőszegi, B. and Rabin, M. (2006), ‘A model of reference-dependent preferences’, *Quarterly Journal of Economics*, **121**, 1133–1165.

Laibson, D. (1997), ‘Golden eggs and hyperbolic discounting’, *Quarterly Journal of Economics*, **112**, 443–478.

Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A.J., Argamon, S.E., Baguley, T., Becker, R.B., Benning, S.D., Bradford, D.E., Buchanan, E.M., Caldwell, A.R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L.J., Collins, G.S., Crook, Z., Cross, E.S., Daniels, S., Danielsson, H., DeBruine, L., Dunleavy, D.J., Earp, B.D., Feist, M.I., Ferrell, J.D., Field, J.G., Fox, N.W., Friesen, A., Gomes, C., Gonzalez-Marquez, M., Grange, J.A., Grieve, A.P., Guggenberger, R., Grist, J., van Harmelen, A.-L., Hasselman, F., Hochard, K.D., Hoffarth, M.R., Holmes, N.P., Ingre, M., Isager, P.M., Isotalus, H.K., Johansson, C., Juszczyk, K., Kenny, D.A., Khalil, A.A., Konat, B., Lao, J., Larsen, E.G., Lodder, G.M.A., Lukavský, J., Madan, C.R., Manheim, D., Martin, S.R., Martin, A.E., Mayo, D.G., McCarthy, R.J., McConway, K., McFarland, C., Nio, A.Q.X., Nilsonne, G., de Oliveira, C.L., de Xivry, J.-J. O., Parsons, S., Pfuhl, G., Quinn, K.A., Sakon, J.J., Saribay, S.A., Schneider, I.K., Selvaraju, M., Sjoerds, Z., Smith, S.G., Smits, T., Spies, J.R., Śreekumar, V., Steltenpohl, C.N., Stenhouse, N., Świątkowski, W., Vadillo, M.A., Van Assen, M.A.L.M., Williams, M.N., Williams, S.E., Williams, D.R., Yarkoni, T., Ziano, I. and Zwaan, R.A. (2018), 'Justify your alpha', *Nature Human Behaviour*, **2** (3), 168–171

Loomes, G. and Pogrebna, G. (2014), 'Measuring individual risk attitudes when preferences are imprecise', *Economic Journal*, **124**, 569–593.

Luckman, A., Donkin, C. and Newell, B.R. (2018), 'Can a single model account for both risky choices and inter-temporal choices? Testing the assumptions underlying models of risky inter-temporal choice', *Psychonomic Bulletin & Review*, **25**, 785–792.

McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L. (2017), Abandon statistical 23 significance. Preprint, arXiv:1709.07588).

Meier, S. and Sprenger, C. (2010), 'Present-biased preferences and credit card borrowing', *American Economic Journal: Applied Economics*, **2**, 193–210.

Odum, A.L. and Rainaud, C.P. (2003), 'Discounting of delayed hypothetical money, alcohol, and food', *Behavioural Processes*, **64**, 305–313.

Peysakhovich, A. and Naecker, J. (2017), 'Using methods from machine learning to evaluate behavioral models of choice under risk and ambiguity', *Journal of Economic Behavior & Organization*, **133**, 373–384.

Popper, K.R. (1959), *The logic of scientific discovery*. Hutchinson, London.

Quiggin, J. (1982), 'A theory of anticipated utility', *Journal of Economic Behavior & Organization*, **3**, 323–343.

Saha, A. (1993), 'Expo-power utility: A 'exible' form for absolute and relative risk aversion', *American Journal of Agricultural Economics*, **75**, 905–913.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics*, **6**, 461–464.

Simon, H. (2007), 'Testability and approximation', in Hausman, D.M. (ed.), *The Philosophy of Economics: An Anthology*, 3rd edn. Cambridge University Press, New York, NY; 179–182.

Stahl, D.O. (2018), 'Assessing the forecast performance of models of choice', *Journal of Behavioral and Experimental Economics*, **73**, 86–92.

Strotz, R.H. (1955), 'Myopia and inconsistency in dynamic utility maximization', *Review of Economic Studies*, **23**, 165–180.

Ubfal, D. (2016), 'How general are time preferences? Eliciting good-speci_c discount rates', *Journal of Development Economics*, **118**, 150–170.

Winer, R.S. (1997), 'Discounting and its impact on durables buying decisions', *Marketing Letters*, **8**, 109–118.

Appendix I

Differences in Data Analysis

Choi et al. (2007)

Choi *et al.* (2007) estimate curvature and disappointment aversion parameters with CRRA utility functions. Their estimation procedure minimises the loss function

$$\sum_{i=1}^{50}\left[\ln\left(\frac{x_1^i}{x_2^i}\right)-f\left(\ln\left(\frac{\overline{x}_2^i}{\overline{x}_1^i}\right);\alpha,\rho,\omega\right)\right]^2,$$

where f is the ratio of demand that arises from a utility function with α and ρ as the parameters, $\overline{x}_1^i$ and $\overline{x}_1^i$ are the maximum values in the budget, and ω is an exogenously chosen value that accounts for consumption ratios of 0 or infinity. They use a similar procedure to estimate parameters from the CARA model, but instead of using the log ratio of consumption as the left-hand-side variable, they use the difference in consumption between the two goods.

Key Differences

1. Instead of using log consumption or consumption differences, the analysis here calculates budget-proportion demands, and minimises sum of squared differences between these budget-proportion demands and the data.
2. As a result of not relying on demand *ratios*, the analysis here does not need the extra parameter ω, which prevents the demand ratio from being 0 or infinite.
3. The analysis from the main text omits a number of subjects due to a variety of concerns, such as low CCEI scores or particular choice patterns. This analysis follows the results from the appendix, which includes the full set of subjects.

Andreoni and Sprenger (2012)

Andreoni and Sprenger (2012) estimate a number of different utility specifications, with both NLLS and Tobit analysis. In their NLLS section, which is the most closely related to what is done here, they calculate the CRRA demand functions from the agent's maximisation problem, and estimate the parameters which minimise the sum of square residuals.

Key Differences

1. The original analysis uses the level of demand for earlier payment as the left-hand-side variable, while the analysis here uses the proportion of the budget devoted to that good. This second approach weights each decision equally, while the former will place higher weight on decisions in which the budget is larger.
2. The analysis here never uses the 'reported' level of background consumption, or ω. Instead, it either assumes that the subjects are choosing as if the background consumption is 0, or estimates it directly.
3. The original analysis estimates the curvature parameter α rather than the parameter ρ. This is just notation, and one can define $\alpha = 1 - \rho$.
4. The original analysis does not estimate preference parameters for a small subset of subjects, either because there was not enough variation in their choices to identify a parameter, because the estimation process did not converge, or because their behaviour exhibited strange choice patterns. The analysis here uses the data from all subjects. If the parameter estimates themselves were a primary objective, the lack of choice variation would be a first-order issue. However, since this paper is about comparing models' predictions, the data can still be used. Regardless, subjects which have little to no variation in their choices are easily predicted by *all* models, so this concern does not drive the results.

Augenblick et al. (2015)

In the section on individual analysis, Augenblick et al. use NLLS on log consumption ratios to get individual parameter estimates for subjects' discounting parameters.

Key Differences

1. Instead of using the log consumption ratio as the left-hand-side variable, the analysis here uses an NLLS estimation with budget shares as the left-hand-side variable.
2. The analysis here uses all monetary delay lengths. The main analysis of the original paper focuses on monetary delay lengths of 3 weeks for easier comparison with the nonparametric tests.
3. The original analysis estimates time preferences over effort for 80 subjects, and a subset of 75 for time preferences over money. For easier comparison, the analysis here focuses only on the 75 subjects for whom there are both kinds of data.
4. The analysis here never uses the 'reported' level of background consumption, or ω. Instead, it either assumes that the subjects are choosing as if the background consumption is 0, or estimates it directly.
5. Here, cost and utility function curvature parameters are estimated for each individual, while in the original work a single curvature parameter is estimated for the entire sample.