On the futility of estimating utility functions: Why the parameters we measure are wrong, and why they do not generalize

Neil Stewart, Emina Canic, Timothy L. Mullett
Warwick Business School, University of Warwick Coventry, CV4 7AL, England, neil.stewart@warwick.ac.uk

We have known for a long time that people's risky choices depart systematically from expected utility theory, and also from related models like prospect theory. But it is still common to use expected utility theory or prospect theory to estimate parameters like risk aversion from sets of risky choices. We have also known for a long time that when parameters are estimated, a systematic departure between the model and the data causes biased parameter estimates. Here we show how the bias in parameter estimation interacts with the set of choices presented to participants. We find that estimates of risk aversion vary greatly between choice sets even though no real differences in risk aversion exist. We find parameters do not generalise at all between choice sets, even when the sets are random draws from a master choice set.

Key words: expected utility theory, prospect theory, bias, risk aversion
History: This revision was compiled January 26, 2020.

1. Introduction

Consider a choice between a sure \$100 and a more risky 50/50 gamble to win \$200 otherwise nothing. There are individual differences in people's choices to questions like this. Most people are risk averse and so prefer the sure \$100. The expected utility EU model is often used to infer the level of risk aversion from a set of choices. In the expected utility theory (von Neumann and Morgenstern 1947) model and related models like prospect theory (Kahneman and Tversky 1979), risk aversion is represented as the curvature of a function $u(\cdot)$ which maps wealth into utility. People are assumed to select the gamble which offers them the highest average, or expected, utility. For the simple choice here, the corresponding expected utilities are $u(100)$ and $\frac{u(0)+u(200)}{2}$. Figure 1 shows a concave utility function where the utility of \$100 is more than half the utility of \$200 and a linear utility function where the utility of \$100 is exactly half the utility of \$200. People with concave utility functions will prefer the sure \$100. People with linear utility functions will be indifferent between gambles with the same expected value, as in the example here. People with convex utility functions will prefer the 50/50 gamble.

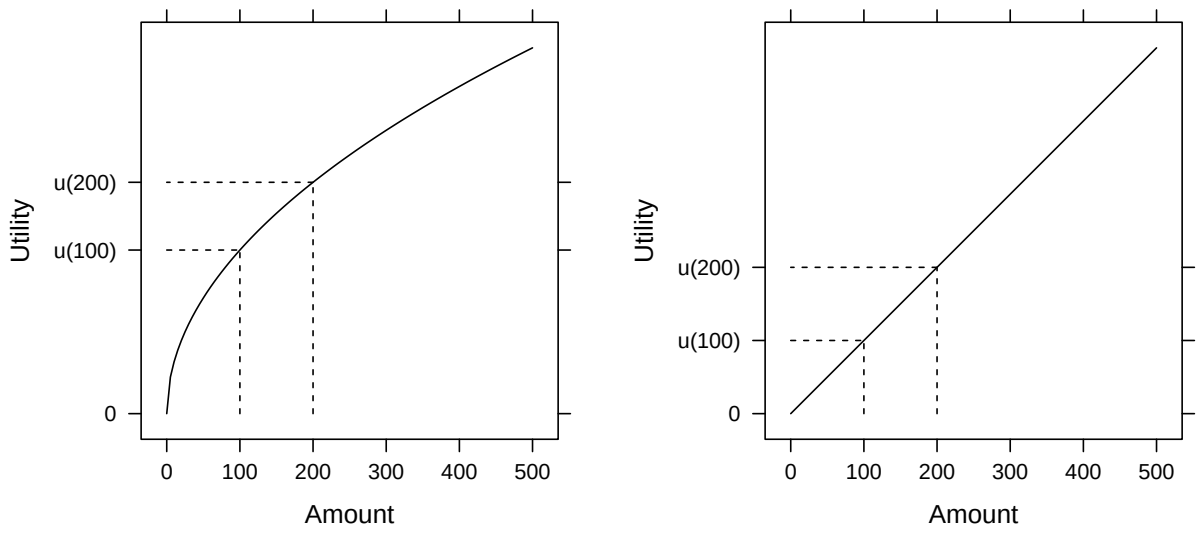

Figure 1 **A concave and a linear utility function. For the concave function the utility of $100 is more than half the utility of $200. For the linear utility function the utility of $100 is exactly half the utility of $200.**

The shape of a person's utility function embodies their preferences over risky options—their level of risk aversion. Risk aversion is considered to be a stable individual difference (see Glöckner and Pachur 2012, Frey et al. 2017, but see Pedroni et al. 2017). Risk aversion is often estimated from choices between simple gambles (Holt and Laury 2002, 2005). Risk aversion is correlated with demographic variables like age and sex (e.g., Croson and Gneezy 2009, Mata et al. 2011, but see Nelson 2015), and possibly psychological variables like intelligence (e.g., Benjamin et al. 2013, Burks et al. 2009, Dohmen et al. 2010, though it may be consistency not risk aversion that is correlated with intelligence Andersson et al. 2013).

In this paper we are going to demonstrate a serious problem in using choice data to estimate utility functions. The risk aversion measure estimated from responses to a set of choices is systematically biased. Because expected utility is the wrong model, its parameters are systematically biased by the systematic variation in the data that the model fails to capture. This is the classic ommited variable bias. In a second, different set of choices risk aversion is also systematically biased. But, in this second set of different choices, the systematic variation in the data that the model fails to capture is different (because the choices are different) and so the bias is different. We are going to show that the variation in the bias induced by different choice sets is large—so large that one cannot generalise the parameters measured with one choice set to other choice sets.

More generally, there are two profound implications: First, the parameter estimates from expected utility theory or prospect theory should not be generalized outside of the context of the choice set used in the estimation: Subtly different choice sets give you very different parameter

estimates. Perhaps this is why risk estimates do not generalize well in economic applications (Einav et al. 2012, Friedman et al. 2014). Perhaps this is why estimates of risk aversion do not appear to generalise from small to large stakes (Bell 2007, Rabin 2000). Second, below we show that the biases in expected utility theory or prospect theory parameters that are introduced when variables are omitted can be large. Practically, this means that the parameter you think is measuring risk aversion can be biased by things that are not risk aversion. We show these biases can be as large as the individual differences between people, and this means, for example, that the person you are identifying as one of the most risk averse in your choice set can actually be one of the least risk averse in another similar but not identical choice set.

2. An Analogy

Consider a simple regression $y = 0 + x + z + \mathcal{N}(0,1)$ where $z = x^2 + \mathcal{N}(0,1)$. Here y is a function of x, and also something which is nonlinearly related to x—that is, z. Figure 2 plots some random data sampled from this model. In these data, y is going to be correlated with x. The curved line is the best fitting quadratic model $y = \beta_0 + \beta_1 x + \beta_2 z + \mathcal{N}(0,\sigma^2)$, and unsurprisingly has a good fit with $\beta_0 = 0$, $\beta_1 = 1$, and $\beta_2 = 1$ because (a) the model we are fitting is the true model—the model that generated the data and (b) these are the values of the coefficients that generated the data. The green and orange lines show the fits that would be recovered if a simple model was estimated using choice sets containing only data where $x > 2$ or only $x < 2$. In this case the simple linear model $y = \beta_0 + \beta_1 x + \mathcal{N}(0,\sigma^2)$ was estimated, with no z term—and this is the wrong model. The green line is estimated on the green data where $x > 2$. For these data, the coefficient $\beta_1 = +7$. The coefficient β_1 is biased away from the value of 1 it took in the model which generated the data. This is the classic omitted variable bias. When we have missed out something correlated with x (in this case z), then the coefficient for x, β_1, must do some of the work that the missing coefficient (β_2) can no longer do. This means that β_1 is not measuring what we think—it is not measuring just the effect of x, but it is also contaminated by measuring the effect of z. Worse still, the bias in β_1 is different for the different ranges. The orange line is estimated on the orange data where $x < 2$. For these data, the coefficient $\beta_1 = 3$. This coefficient is also biased away from the value of 1 it took in the model which generated the data. And this coefficient does not match the value of 7 estimated from the other half of the data when $x > 2$. This is because of the nonlinear relationship between x and the missing z. This means that across different data sets (here different ranges of x), β_1 is differently biased. The omitted variable problem is interacting with the data set used for estimation to create different biases in different data sets.

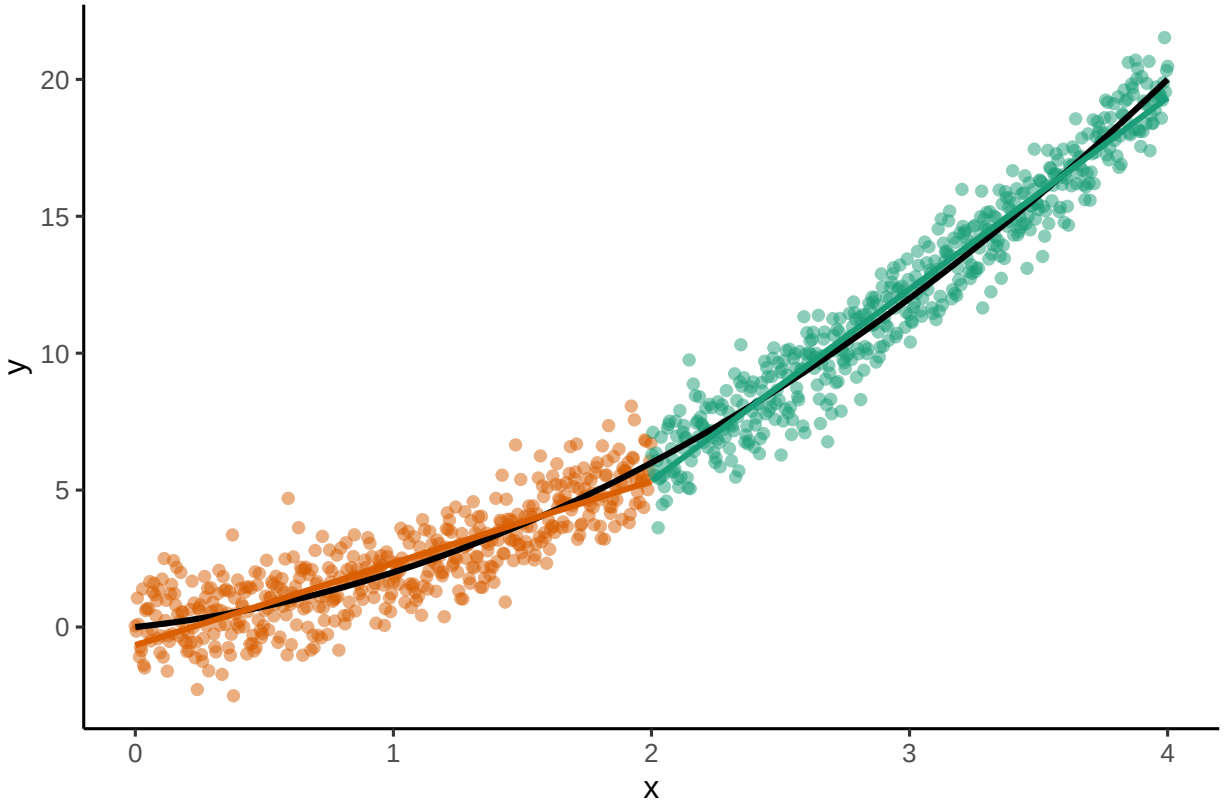

Figure 2 $y = 0 + x + z + \sim \mathcal{N}(0,1)$**, where** $z = x^2 + \mathcal{N}(0,1)$**. The dots are random samples from this model. The curved line is the best fitting regression with** x **and** z **terms. The orange and green lines are fits of a regression with just** x **for either** $x < 2$ **or** $x > 2$**.**

3. Overview

We now present three examples of the futility of estimating utility functions from choice data. The first is, by intention, somewhat trivial. We use the common ratio effect to illustrate that risk aversion estimated in an expected utility model is systematically biased in different choice sets. That is, the risk aversion measured by expected utility does not generalise across choice sets. We show this in simulation and also in a data set of real choices from Glöckner and Pachur (2012).

In our second example, we revisit Stewart et al.'s (2015) demonstration that the utility functions estimated from choice data are a property not of the preferences of the individual making the choices but of the experimenter's selection of choices used to measure the individual's preferences. While Stewart et al.'s results are sound, they should be attributed to the futility of estimating utility, and not, as Stewart et al. concluded, to the decision by sampling theory (Stewart et al. 2006).

Our final example extends our analysis of Glöckner and Pachur's (2012) data and shows that the problem we are illustrating is very general. We recover expected utility and prospect theory estimates of risk aversion from thousands of separate random samples of the Glöckner and Pachur

choices. We show that the recovered risk aversion, and other parameters, differs systematically over the summary properties of the random samples, and, most disturbingly, that the reliability of recovered parameters across samples of choices is extremely poor. For example, we show that the rank of a participant's risk aversion in one subset of the choices tells you almost nothing about their rank in another subset of the choices. This third example demonstrates that the futility of estimating utility is not limited to expected utility, but is a more general problem that affects prospect theory (and, presumably, other models), and also that the effect is not specific to carefully selected sets of choices, but applies much more generally.

Data and code for the three examples can be found at https://github.com/neil-stewart/futility_of_utility_code_and_data.

4. Expected Utility and Cumulative Prospect Theory Models

Before we come to the examples, we introduce a reasonably standard formulation of an expected utility model where utility is constrained to be a power of money. This power assumption is very common in fitting the expected utility model and related models like prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992) and the transfer-of-attention-exchange model (Birnbaum 2008) to choices between gambles (see Stott 2006 for a review of functional forms in estimating prospect theory).

We apply the model to simple choices between a relatively risky gamble G_{risky} and a relatively safe gamble G_{safe}, each of the form "p chance of x_1 otherwise x_2" where $x_1 > x_2 > 0$.

The utility (or subjective value) of a sum of money x is given by a power function

$$u(x) = x^{\alpha} \tag{1}$$

where $\alpha > 0$. When $\alpha = 1$ the utility function is linear. When $\alpha < 1$ the utility function will be concave and produce risk averse choices and when $\alpha > 1$ the utility function will be convex and produce risk seeking choices.

The expected utility of a gamble $G = \{p_1,\ x_1;\ 1 - p_1,\ x_2\}$ is given by weighting utilities by probabilities and summing

$$U(G) = p\ u(x_1) + (1 - p)\ u(x_2) \tag{2}$$

To make probabilistic predictions we use the Luce choice rule. The probability of choosing the safer gamble is given by

$$Prob(Safe) = \frac{U(G_{safe})^{\phi}}{U(G_{safe})^{\phi} + U(G_{risky})^{\phi}} \tag{3}$$

In this model, the sensitivity parameter ϕ controls the degree of determinism in the model, such that when ϕ is zero people choose randomly, when $\phi = 1$ choice probabilities follow the ratio of the

expected utilities, and when $\phi > 1$ people have a strong tendency to choose the gamble with the higher expected utility, even if it is only slightly better.

We also use a standard form of cumulative prospect theory (CPT) applied to gambles in the domain of gains. To apply CPT to the domain of gains only, we do not need to be concerned with loss aversion and the differences in the curvature of the value function applied to gains and losses. Thus we have the same form for prospect theory value as we did for utility in the expected utility model.

$$v(x) = x^{\alpha} \tag{4}$$

Where we need to discriminate between α from expected utility and α from prospect theory we will add subscripts.

Prospect theory also transforms gamble probabilities into decision weights. We use the Tversky-Kahneman form (see Stott 2006 for other forms).

$$w(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{\frac{1}{\gamma}}} \tag{5}$$

where γ controls the curvature of the weighting function. When $\gamma = 1$ the model reduces to expected utility theory. When $\gamma > 1$, probability weighting takes the typical inverse-S shape where small probabilities are overweighted and large probabilities are underweighted.

The cumulative prospect theory value of gamble G is

$$V(G) = w(p)\ v(x_1) + (1 - w(p))\ v(x_2) \tag{6}$$

Again, to make probabilistic predictions we use the Luce choice rule.

5. Example 1: The Common Ratio Effect

In the common ratio effect (Allais 1953 and see, for example, Kahneman and Tversky 1979 and Starmer 2000), people are risk averse when making choices between lotteries with moderate to large probabilities, but become risk seeking when the choices are between lotteries with small probabilities. For example people are risk averse when choosing between a 80% chance of 4,000 otherwise nothing vs a sure 3,000, selecting the sure 3,000. But when probabilities of winning are scaled down by a factor of four the model preference is reversed. Continuing the example, people prefer a 20% chance of 4,000 otherwise nothing to a 25% chance of 3,000 otherwise nothing. The common ratio effect is one of the phenomena that motivated the development of non-linear weighting of probabilities in, for example, subjective expected utility and prospect theory. We take this well-established effect to illustrate our general argument in this paper. So while the common ratio effect is typically understood as the empirical evidence for a non-linear probability weighting

function, we are taking a different perspective and using the common ratio effect as a demonstration that risk aversion estimated by expected utility is differently systematically biased in different choice sets.

5.1. A Simple Simulation

Here we are going to generate data from prospect theory and then fit expected utility theory. Because expected utility theory is not the data-generating model, the parameter estimates from expected utility are going to be systematically biased. We will show that the systematic bias varies over the set of choices for estimation. Specifically, we will show that, when the set of choices used has large probabilities, expected utility risk aversion is higher than when the set of choices used has small probabilities. The explanation is due to the probability weighting function in prospect theory. The weighting function is flatter for small probabilities, which means that people differentiate between small probabilities less and, thus, take more risk—as they are more willing to trade off a reduction in probability for an increase in amount. The weighting function is steeper for large probabilities, which means that people differentiate more between large probabilities and, thus, take less risk—as they are less willing to trade off a reduction in probability for an increase in amount. In short, expected utility parameter estimates are biased because expected utility does not capture the non-linear weighting of probabilities, and the bias this omission creates is different in choice sets with different distributions of probability.

First we created two sets of probabilities: one of Small Probabilities (0.01, 0.02, 0.05, 0.10, 0.15, 0.20) and one of Large Probabilities that are five times larger (0.05, 0.10, 0.25, 0.50, 0.75, 1.00). We used the set of amounts (10, 20, 50, 100, 200, 500). We made all gambles of the form "p chance of x otherwise y" where $(x > y)$. We crossed the set of gambles with themselves to make all possible choices, dropping choices where one gamble was the same as or stochastically dominated the other. This created a set of 3,006 choices.

We generated choice data for 500 simulated CPT participants. We used the parameters $\alpha = 1$ for risk aversion, $\gamma = 0.5$ for probability weighting, and $\phi = 0.1$ for sensitivity for all participants. For each participant we sampled 200 choices at random from the set of 3,006 and generated Bernoulli 0/1 responses from the CPT choice probabilities.

Figure 3 shows the median parameters recovered by fitting expected utility to the choices for each of the 500 simulated participants. We conducted the exercise separately for (a) each simulated participant's 200 choices, (b) only those choices where the probabilities from each gamble were from the Small set, and (c) only those choices where the probabilities from each gamble were from the Large set. The median α is 0.73 95% CI[0.69–0.77] when using only Large Probability choices and 1.23 95% CI[1.15–1.32] when using only Small Probability choices.

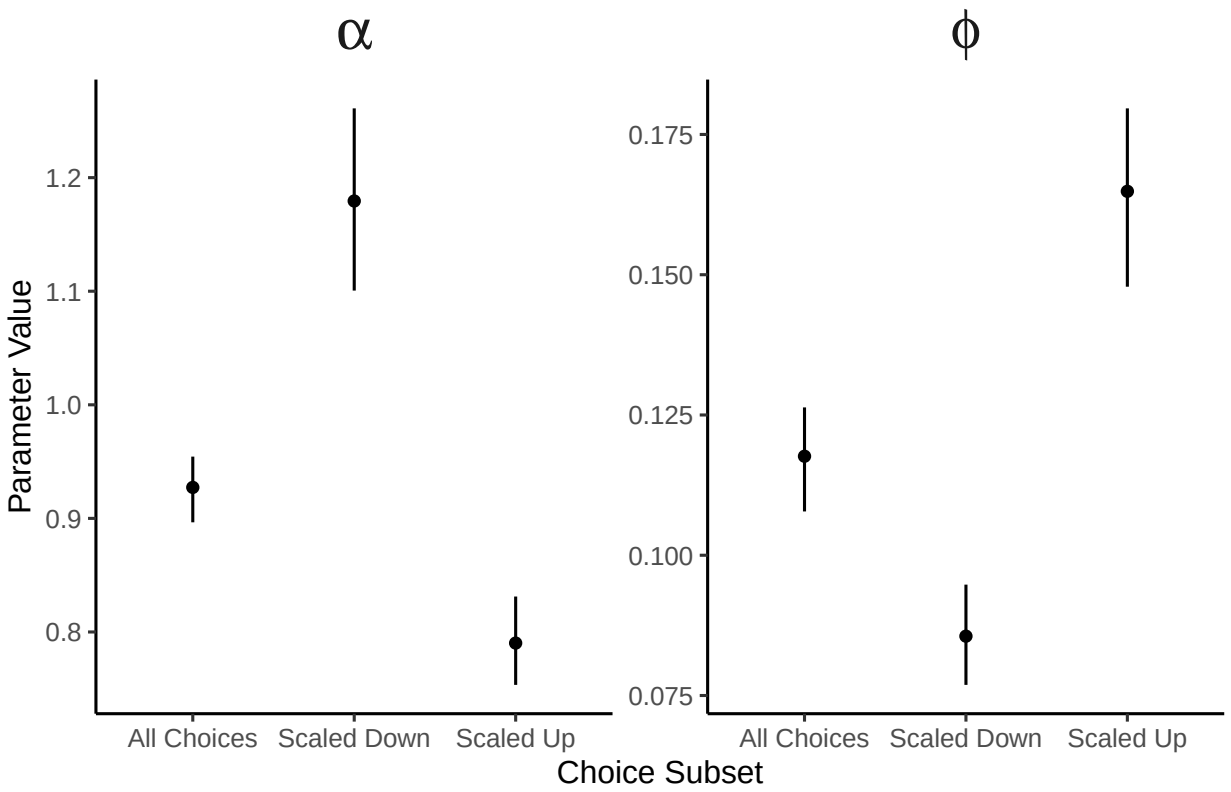

Figure 3 Median risk aversion (left) and sensitivity (right) recovered from the expected utility model fitted to fake data generated from CPT. Error bars span the bootstrapped 95% confidence intervals.

This result—risk aversion with moderate-to-large probability gambles and risk-seeking with small probability gambles—is exactly what we'd expect given that the weighting function was included in prospect theory to capture the common ratio effect. But it demonstrates that the parameters estimated from expected utility, which is, in this case, not the data-generating model, are biased, and are biased differently in different choice sets.

5.2. Fits to Real Choices

We now demonstrate, using choice data from Glöckner and Pachur (2012), that the selection of choices used to estimate risk aversion does indeed systematically bias the estimation of risk aversion. We will split the choice set into subsets based on the size of the probabilities in the choices, and show that, because of the common ratio effect, the risk aversion estimated varies across these sets.

Glöckner and Pachur (2012) used choices between two-outcome gambles of the form "p chance of x otherwise y". Their set comprised randomly generated choices from Rieskamp (2008), problems designed to separate CPT and the priority heuristic from Glöckner and Betsch (2008), choice problems designed to measure risk aversion from Holt and Laury (2002), and choice problems designed to measure loss aversion from Gäechter et al. (2007). There are several benefits of using

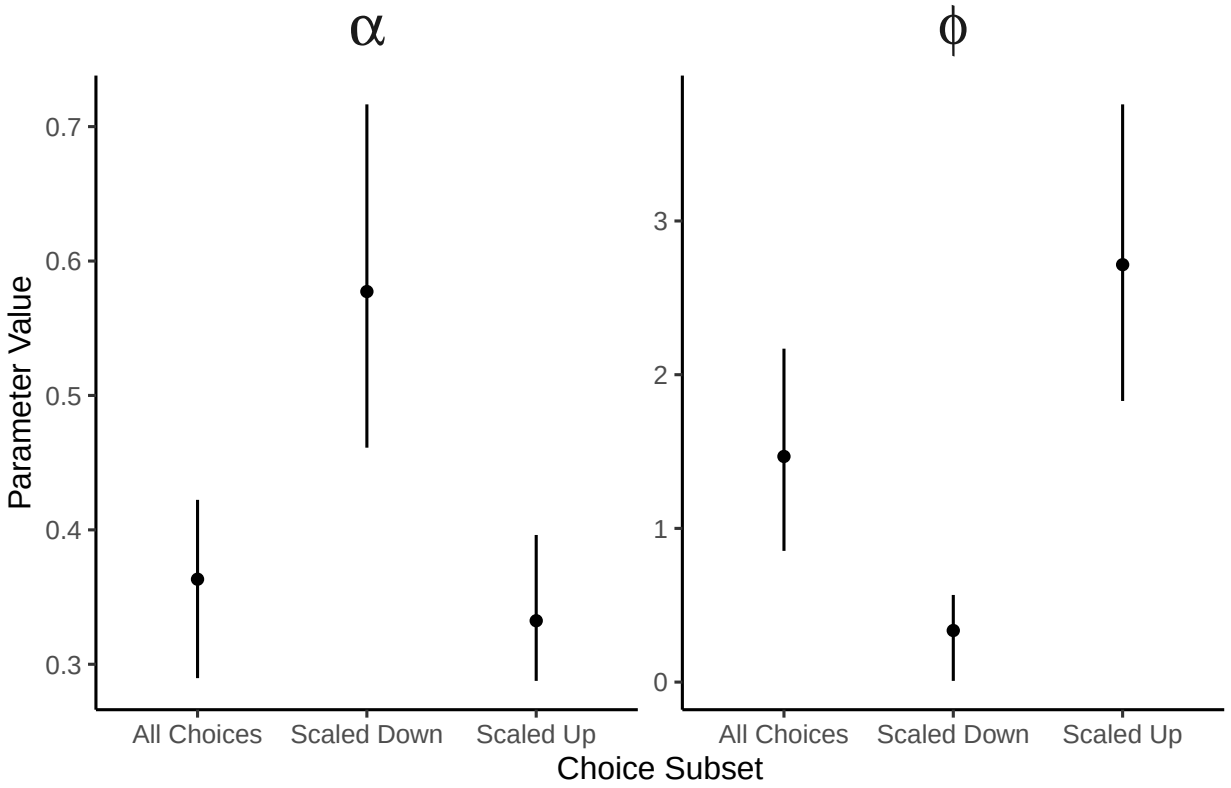

Figure 4 Median risk aversion (left) and sensitivity (right) recovered from the expected utility model fitted to fake data generated from prospect theory. Error bars span the bootstrapped 95% confidence intervals.

this dataset. One is that a central purpose of the original paper was to measure test-retest reliability of CPT parameters within an individual. Furthermore, the dataset has also been used in a number of subsequent reanalyses and meta-analyses, and the choice set has been shown to have good power for parameter identification (Broomel and Bhatia 2014). We dropped data from 5 of the 66 participants who completed half of the choices in the set twice instead of completing the whole set once and 2 participants who did not complete all of the choices. From the total set of 237 unique choices we took only the 109 from the domain of gains.

Figure 4 shows the median parameters recovered by fitting expected utility to the choices. We conducted the exercise separately for (a) each participant's 200 choices, (b) only those choices where the probabilities of winning from each gamble were less than one half (the Small set), and (c) only those choices where the probabilities of winning from each gamble were greater than one half (the Large set). The median α was 0.33 95% CI[0.29–0 0.40] for the Large Probability choice set and 0.58 95% CI[0.46–0.71] for the Small Probability choice set. This replicates the pattern seen in the simulation using real choice data (although in the Glöckner and Pachur the overall level of risk aversion is higher than in the simulation). Depending on which subset of the choices we use

from Glöckner and Pachur, we obtain quite different levels of risk aversion. That is, the level of risk aversion recovered is a property of the choice set used and not (just) of the preferences of the individual.

6. Example 2: Revisiting Stewart et al. (2015)'s Malleable Utility Functions

In our second example we revisit the demonstration of malleable utility functions from Stewart et al. (2015). Stewart et al. demonstrated that the utility functions revealed from people's choices depend heavily upon the set of choices used to elicit them. This is most troublesome for the view that the shape of the utility function is a stable property of the individual.

In Stewart et al.'s experiment half of the participants received choices from the Positive Choice Set where the distribution of amounts £10, £20, £50, £100, £200, and £500 on offer was positively skewed. Half of the participants received choices from the Uniform Choice Set where the distribution of amounts £100, £200, £300, £400 and £500 on offer was uniformly distributed. Although people were randomly assigned to choice sets, the revealed utility functions differ greatly—which means that the utility functions are, to a large part, a property of the choice set used to elicit them and not just of the risk preference of the individuals. Figure 5 shows the utility functions revealed from the Positive and Uniform choices from a similar experiment by Canic and Stewart reported in Alempaki et al. (2019) (see Appendix A for method).

This finding is robust, as demonstrated by a meta analysis of 13 experiments including this one (Alempaki et al. 2019). Stewart et al. (2015) attribute this result to the decision by sampling model (Stewart et al. 2006, Stewart 2009). According to decision by sampling, people make comparisons between attribute values (here, amounts of money and probabilities) in their memory. They accumulate the number of comparisons that favor each gamble until a threshold is reached and a choice is made. The probability of a favorable comparison is a function of the number of attribute values in memory that are smaller. This means that £200, say, is more likely to win when a person's memory is full of amounts available in the Positive Choice Set because there are more amounts smaller than £200 (specifically 4/5 are smaller: £10, £20, £50, £100, but not £500). And £200 is less likely to win when a person's memory is full of amounts available in the Uniform Choice Set (specifically 1/4 are smaller: £100, but not £300, £400 or £500). More generally, people will behave as if a sum of money has a higher utility the more smaller sums of money they have in mind. Thus the utility functions in Figure 5 should look like cumulative density functions for the distribution of amounts of money, because the cumulative density function gives the fraction of amounts in the choice set that are smaller than a given value.

Here we show that attributing the revealing of different utility functions from different choice sets to decision by sampling was a mistake. Stewart et al. (2015)'s mistake was to use the expected utility

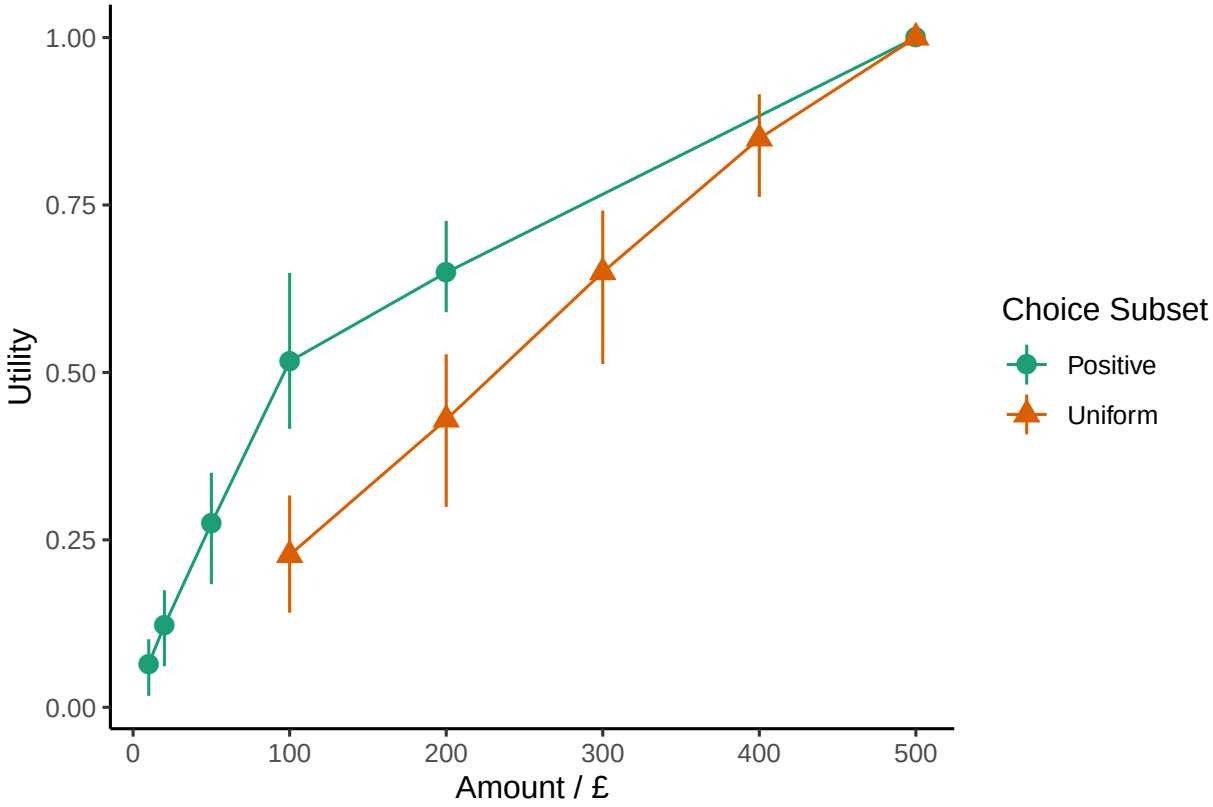

Figure 5 **Revealed utility functions estimated from an unpublished experiment by Canic and Stewart (see Appendix A). Error bars are boostrapped 95% CIs. Labels indicate the choice sets, which were manipulated within participants.**

model to reveal utility functions from choice data. Because people's choices depart systematically from the expected utility model, the parameters estimated in this model are biased. Here we will show how this systematic bias varies as a function of the set of choices used—and that this is the source of the effects that Stewart et al. report. Parenthetically, if you are concerned about the fate of the decision by sampling model, let us reassure you that there are other sources of evidence to support the model unaffected by this problem (see Canic 2016, Noguchi and Stewart 2018, Stewart 2009, Stewart and Simpson 2008, Mullett and Tunney 2013).

In the Canic and Stewart experiment, each participant completed choices from the Positive Choice Set and the Uniform Choice Set, mixed together in a random order (cf. the original experiments where choice set was a between-participants manipulation). It would be very hard for participants mentally to partition choices into the two different sets. They have no motivation to do this and it would be implausible to claim that they spontaneously did so. Thus people could not behave, continuing the earlier example, as if £200 has a higher utility in choices from the Positive Condition compared to the Uniform Condition, because there is no way for them to separate out

the sums of money seen earlier into the Positive and Uniform distributions. However we, as experimenters, did retrospectively separate the mixed choices into the Positive and Uniform Choice Sets, and we estimated utility functions separately using these subsets. Because the choice sets were mixed together, participants would have had all of the amounts in the experiment in mind at any one time, so the decision by sampling account of any difference in the utility functions across these subsets is no longer viable.

Figure 5 shows the utility functions revealed by retrospectively separating the data into Positive and Uniform Choice Sets, and fitting a variant of the expected utility model in which the utility of each sum of money is a separate free parameter (which contains Equation 1 as a special case). We did this separately for each participant, and Figure 5 shows the bootstrapped median utility functions. We see quite different median utility functions estimated for each choice set.

To understand why different utility functions are revealed from different subsets of choices, even though they are from the same individuals, it is useful to consider the stochastic expected utility model as a logistic regression, which is possible for Canic and Stewart's simple choices with exactly one non-zero outcome Stewart et al. (2015). For convenience let us label the riskier gamble as "p chance of x otherwise 0" and the safer gamble as "q chance of y otherwise 0" where $p < q$ and $x > y$. To make probabilistic predictions we again use the Luce choice rule. The probability of choosing the safer gamble is given by

$$Prob(Safe) = \frac{bias(q\,y^{\alpha})^{\phi}}{bias(q\,y^{\alpha})^{\phi} + (p\,x^{\alpha})^{\phi}} \tag{7}$$

This is as Equation 3, but we have substituted for G_{safe} and G_{risky} and added a additional *bias* parameter. The *bias* parameter controls a tendency to choose the safer gamble independent of the actual probabilities and amounts on offer. The need for a *bias* parameter become obvious when Equation 7 is rewritten as a logistic regression (see Stewart et al. 2015's appendices). Rearranging Equation 7 gives

$$\log \frac{Prob(Safe)}{1 - Prob(Safe)} = \log(bias) + \phi\alpha \log(y/x) + \phi \log(q/p) \tag{8}$$

The log odds of a safe choice are a linear function of the log of the ratio of amounts $\log(y/x)$ and the log of the ratio of probabilities $\log(q/p)$. This linear-in-log-odds form is convenient because we can use off-the-shelf logistic regressions to estimate parameters by maximum likelihood and because it is easy to illustrate the effect of model parameters. In particular it shows us that the slope on a plot of log odds of a safe choice against $\log(y/x)$ is given by $\phi\alpha$.

Figure 6 shows that the underlying cause of the difference in utility functions is very similar to that laid out in the opening analogy in Section 2. The data show that the log odds of a safe

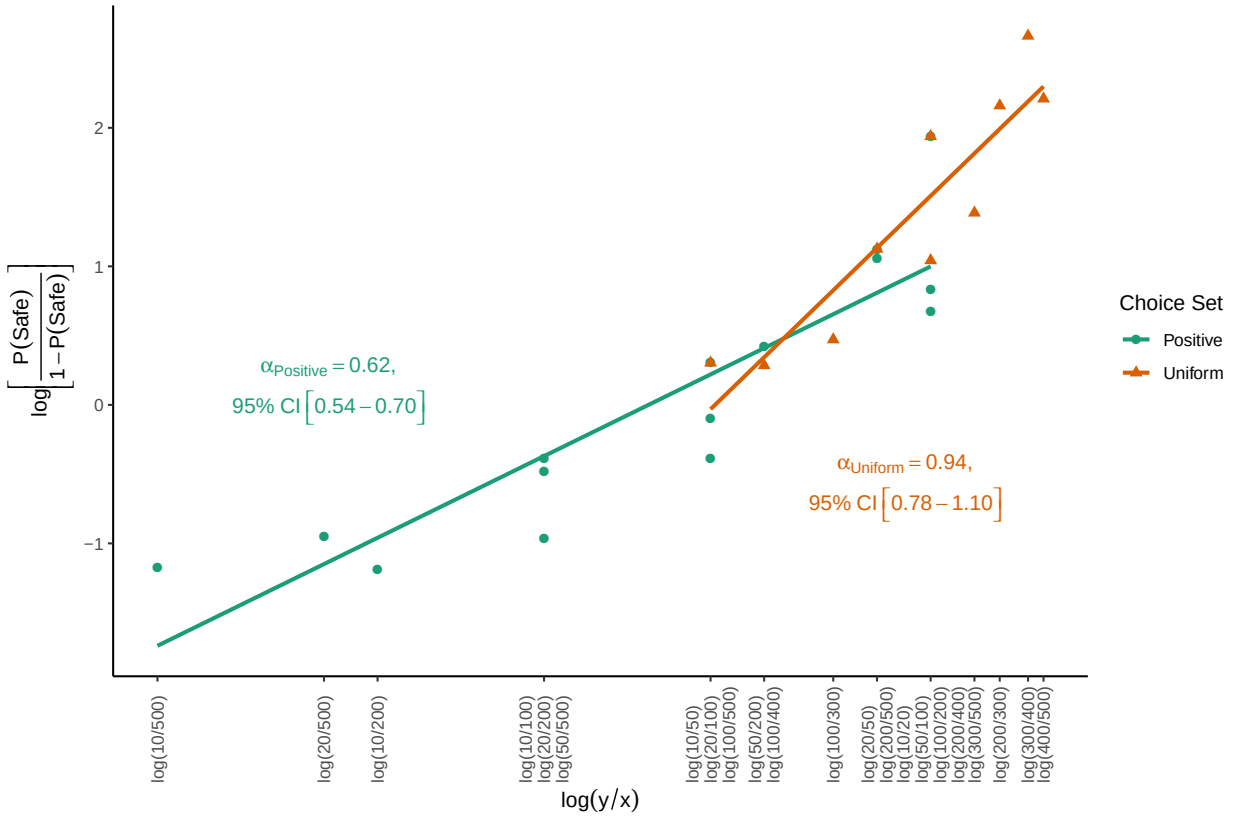

Figure 6 Fits of expected utility to Positive and Uniform choice sets. The log odds of a safe choice are plotted as a function of $\log(y/x)$. The factorial combination of $\log(q/p)$ ratios with $\log(y/x)$ ratios allows us to average the log odds of a safe choice over choices with different probabilities.

choice increases as a function of $\log(y/x)$. Although Equation 8 shows that this increase should be linear under expected utility, the increase is non-linear in the data in Figure 6 with the slope flatter on the left of the plot where y is small compared to x and but steeper on the right of the plot where y is larger compared to x. We describe this as a prize-similarity effect: people are less risk averse when the two prizes are similar. This means that $\phi\alpha$, and thus α, is smaller when y is smaller compared to x and larger when y is large compared to x. Because the Positive Choice Set comprises more choices where y is small compared to x, this means that α from fits to the Positive Choice Set is smaller—corresponding to more risk aversion and a concave utility function. And because the Uniform Choice Set comprises more choices where y is large compared to x, this means that α from fits to the Uniform Choice Set is larger—corresponding to less risk aversion and a more linear utility function.

So Stewart et al.'s result shows not evidence for decision by sampling, but instead that when a model systematically departs from the data, its parameters are biased—and that this bias is different in different choice sets. While Stewart et al. were wrong in concluding that the effect is

evidence for decision by sampling, they were correct in concluding that the level of risk aversion measured is a property of the choice set used to elicit it.

Thus far we have seen how two systematic departures from expected utility—the common ratio effect in Section 5 and a win-similarity effect in Section 6—create a large bias when estimating risk aversion from expected utility and, additionally, that this large bias is different in different choice sets. There are many other behavioural effects departing from expected utility (see Starmer 2000 and Bhatia et al. 2016 for many tens of effects). In the next section, rather than focus on a particular behavioural effect, we make a systematic exploration of parameter values as a function of many choice set properties. We also demonstrate that specially constructed choice sets are not required to see large differences in parameters across choice sets—randomly sampled choice sets from a larger choice set are sufficient.

Our next example also extends our argument from expected utility theory to prospect theory. There are known departures from prospect theory, such as Birnbaum's (2008) paradoxes, which suggests that estimating prospect theory parameters might also be futile—at least in the sense that prospect theory parameters might also generalize rather badly.

7. Example 3: It Is Worse Than We Though: Estimating Prospect Theory Parameters is Futile Because They Do Not Generalise Even Across Random Splits of a Choice Set

It is possible that in the examples thus far, we have stumbled upon rare, specific effects where parameter recovery is particularly unreliable. Perhaps these issues are limited only to specific choice phenomena that are not captured by simple models. Perhaps they are only apparent when using narrow stimuli sets that have been specifically designed to induce them. It may also be that parameter stability increases with more complex models that capture more properties of choice behaviour. In this section, we take the data from Glöckner and Pachur (2012) and refit choices using both expected utility and prospect theory. Glöckner and Pachur show that prospect theory parameter estimates are stable over a time period of one week when they use the same choices to estimate parameters at each time point. Here we examine the reliability of parameter estimates across random partitions of the choice set. While Glöckner and Pachur find stability in parameters over time, we find almost no stability over choice sets.

To examine the effect of changing the choices used in an experiment, we take random subsamples of the overall choice set and estimate EU and CPT for each subject. Subsamples are generated by randomly splitting the 108 unique choice items in two halves, with 54 choices in each. Note that due to the design of the blocks in Glöckner and Pachur, subjects answered the majority of these

unique choices twice, increasing our statistical power for parameter identification. A total of 10,000 splits of the choice set were used, meaning expected utility and prospect theory fits were calculated for a total of 20,000 subsets. Thus, rather than identifying a property of the choice set that we believe may result in different parameter fits and then selecting the two extreme examples on this dimension, this random sampling approach allows us to identify whether parameter estimates vary due to random variation in choice selection.

7.1. Test-Retest Reliability With Different Choice Subsets

One of the most fundamental questions for future research is whether model fitting will produce reliable estimates of a subjects model parameters across different choice sets. It is possible that even if changes to the choice set result in changes to the absolute estimates of parameter values, the relative estimates between subjects will remain stable. That is, the average estimate of a parameter may be different in different choice sets, but if the estimate is affected similarly for each subject it could still be possible to identify individuals who are relatively risk averse or relatively risk seeking; an individual at the 25th percentile in one choice set would be close to the 25th percentile in another. In this section we use cross validation to examine the stability across choice sets of individuals relative position in the distribution of risk aversion.

In fitting expected utility and prospect theory models to responses in this task, we are assuming two sources of noise in the estimate of an individuals rank position in risk aversion. One is the effect of changing the choice set. The other is the natural stochasticity in preference. The latter is incorporated into expected utility and prospect theory as the sensitivity parameter ϕ. We take a cross validation approach, which allows us to identify the effect of stochasticity, and then see whether there is additional effect of changing the choice set.

First we provide a baseline for what proportion of the noise is explained by the models stochasticity alone. To do so, the parameter estimates from the first half of the random split were estimated. These were then used to generate "within-half" choice predictions for those same choices and same subjects. These predictions were inherently noisy, as the probabilistic predictions in the model were transformed into binary choices (i.e. choices with an 0.7 prediction would resolve to 1 70% of the time and to 0 30% of the time). We then estimated parameters using these generated choices. By comparing the parameters used to generate the simulated choice data with the parameters recovered from fitting the simulated choice data, we can estimate how much parameters should vary because of stochasticity in responding alone. By comparing the distribution of these within-half recovered parameters with the distribution of parameters estimated from subjects responses in the other subset half, we can identify whether there is additional impact of changing the choice set.

To illustrate the effect, a number of points are identified in the distribution of alpha parameters: values of α that are the 5th, 25th, 50th, 75th, and 95th percentile. For each split of the choice set,

the distribution of alphas in the first half is used to identify the individual in that position. For that individual, their rank position is then also found in the distribution of alpha values in the second half of the data, and in the within-half predictions.

The top panel in Figure 7 plots in orange the distribution of the rank of a participant in the second half of choices for those at a given exact percentile in the first half of choices. For example, the top middle panel shows the distribution of the rank in the second half of choices for participants who were at exactly the 50th percentile in the first half of choices. If estimation of α were consistent, there should be a single spike at the 50th percentile. Instead we see that the distribution is roughly uniform over the range 0–100%. Knowing that a person is that the 50th percentile in the first half of choices tells you almost nothing about their ranking in the second half of choices. The green distribution shows the spread over ranks that would be expected from stochasticity alone—in all cases the spread from the first to second half of choices exceeds that which would be expected from stochasticity alone. While consistency is more stable for individuals who have extreme values of α estimated from the first half of choices, but there is still a significant proportion of subsamples where an individual will move from the most risk averse decile to the least risk averse decile, or vice versa.

The futility of estimating utility is seen in prospect theory as well as expected utility theory. The bottom panel in Figure 7 is the corresponding plot for prospect theory estimations of α. Consistency between α estimated across random splits of the choices is even worse than for expected utility.

7.2. How Choice Set Properties Correlate With Parameter Changes

The results above demonstrate that the variance in parameter estimates is significant, and large compared to individual differences and stochasticity captured by the models themselves. Here we examine whether we can identify reliable relationships between variation in choice set properties, and the variation in parameter estimates. To do so would demonstrate that this additional variance is due to the random selection of choices, even when the overall choice set is balanced.

Here we apply a very general approach, identifying a selection of simple measures on which the splits can vary. For each measure, the score on each individual choice was calculated, then the mean was taken across the choices within each half. The measures are: the maximum single payout available within the choice, the minimum single payout, the maximum probability, the probability of receiving the largest payout, the mean of the two highest payouts, the mean of the two lowest payouts, the mean of all payouts, the range from highest to lowest payout, the difference in expected value between the two options, the mean of the two highest probabilities, and the range of probabilities.

Table 1 shows the beta coefficients from individual regression analyses where each of the measures described above is used to predict the median estimate of α and ϕ for expected utility and,

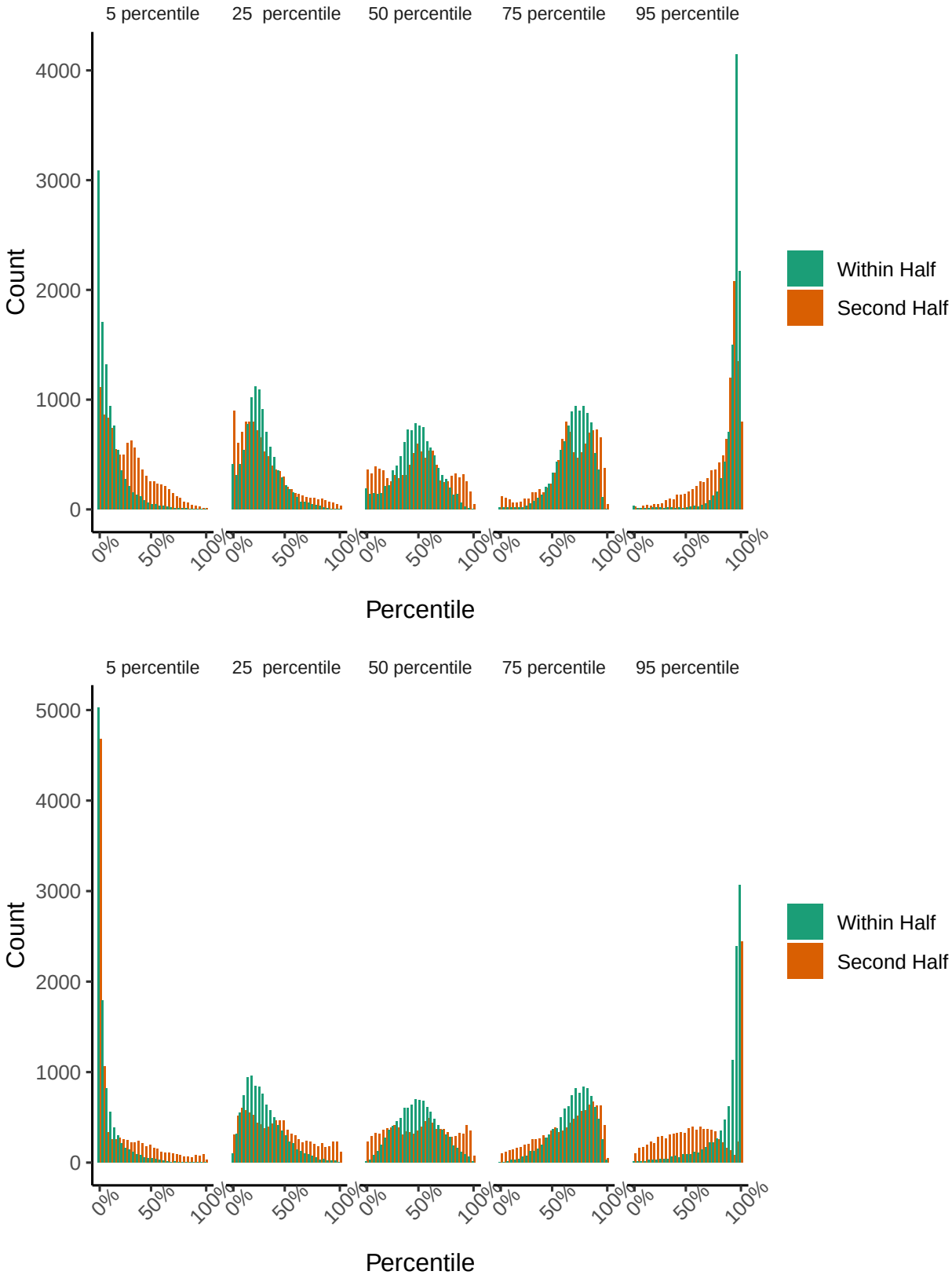

Figure 7 **Generalisation of α between random splits of a choice set for expected utility (top) and prospect theory (bottom)**

Choice set property	CPT α	CPT γ	CPT ϕ	EU α	EU ϕ
Max single payout	0.026***	−0.001	−0.078***	0.013***	−0.001***
Min single payout	0.016***	0.003*	−0.009*	0.013***	0.001***
Max single probability	0.206***	−0.175***	−0.445***	−0.113***	−0.010***
Probability of biggest payout	−0.169***	−0.040***	0.268***	−0.057***	−0.007***
Mean of two highest payouts	0.018***	0.015***	−0.061***	0.036***	−0.001***
Mean of two lowest payouts	0.018***	0.006***	−0.025***	0.012***	0.001***
Mean of all payouts	0.024***	0.014***	−0.057***	0.032***	0.000***
Mean range of payouts within choice	0.007***	−0.002	−0.044***	0.001	−0.001***
Expected Value Difference	0.011***	0.027***	−0.006	0.047***	0.002***
Mean of two highest probabilities	0.300***	−0.293***	−0.636***	−0.170***	−0.010***
Mean range of probs within choice	0.103***	−0.087***	−0.223***	−0.056***	−0.005***

Note: p values are indicated as: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

Table 1 Regressions predicting parameter values from choice set properties.

additionally, γ for prospect theory. The regression analysis is used to provide a measure of both the statistical reliability of the relationship and the size of the effect (i.e. the average amount of change in parameter for a given change in measure for the stimuli set). To facilitate interpretation, predictors are scaled such that the regression weight represents the change in parameter estimation for a change of €100 for value based predictors, and 100% for probability. Median estimates of parameter values are used to limit the impact of the minority of fits that failed to converge during maximum likelihood estimation as well as minimising the effect of skewed distributions and other outliers.

Table 1 shows many properties of the choice set have a statistically significant relationship with the estimated model parameters. Notably, those with the largest effect differ between expected utility and prospect theory estimates. For example, the value of the maximum single probability within a choice decreases α estimates substantially for expected utility but increases α estimates substantially for prospect theory. This difference must be attributed to the probability weighting function in prospect theory—γ which itself shows a significant relationship with the value of the maximum probability, but in the opposite direction to α. The effect of properties of the payout has the same effect on α in expected utility and prospect theory, though the magnitude of the effect is markedly different. For example the mean range of payouts within choices has a significant relationship with α for prospect theory but not expected utility.

Crucially, what this result shows is that both models are significantly affected by many properties of the choice set. In the simpler examples of the previous sections, it may have seemed possible that a more complete model would solve these parameter estimation issues. However, this shows that the effect is a general one, and that even the more complex prospect theory is susceptible. Arguably, it is even more susceptible, as the additional parameters mean is it more susceptible to overfitting. An issue which could explain why CPT performs worse than EU in Figure 7.

	CPT α	CPT γ	CPT ϕ	EU α	EU ϕ
CPT α					
CPT γ	-0.34^{***}				
CPT ϕ	-0.79^{***}	0.35^{***}			
EU α	0.18^{***}	0.23^{***}	-0.13^{***}		
EU ϕ	0.07^{***}	0.14^{***}	0.07^{***}	0.37^{***}	

Note: p values are indicated as: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

Table 2 Correlations between parameter across random splits of the choice set

If the changes in parameter estimates are due to model mis-specification then it is likely that in finding the best fitting parameter values, the parameters must trade off against each other. For example in the experiment in Figure 2, the slope and intercept terms would trade off as they attempt to find the best fitting straight line to different sections of a curvilinear relationship. As a test of this parameter trade-off, median values were taken for each of the parameters in each subsample, and correlation analyses performed between the different parameters.

Table 2 show the correlations between parameters over splits of the choice set. The largest of these is a correlation of $r = -.79$, $p < 0.001$, between α and ϕ in prospect theory (see Stewart et al. 2018 for an explanation). A correlation that one may hope to be strong is between α from expected utility and α from prospect theory: Since both parameters are measuring risk aversion, one would assume that changes in one will be closely mirrored by changes in the other. However, despite being statistically significant, the magnitude of the correlation is small ($r = .18$), showing just 3.2% of shared variance explained. This indicates that changes in choice set affect these models quite differently, again showing that the models are each mis-specified in a different way.

8. Discussion

Here we have shown that, because choice data systematically depart from expected utility theory, the estimates of the risk aversion parameter from that theory are systematically biased. This is the well-known omitted variable problem. Further we have shown that this bias is different in different choice sets. And this is why estimating utility is futile. The bias is large and differs across choice sets. This means that risk aversion estimated in one choice set does not generalise to other choice sets. We have shown this for the well-understood common ratio effect in two carefully designed choice sets. We have also shown this to be true for random choice sets drawn from a single larger choice set. Finding that risk aversion generalises so badly even between choice sets randomly drawn from the same population of choices demonstrates that estimating utility from choices between lotteries is futile. What use is an individual difference measure that does not even generalise to an extremely similar context? What use is a theory that does not make generalisable predictions?

8.1. Stewart, Reimers, and Harris's (2015) Experiments Are Not Evidence For Decision by Sampling

Our results have implications for Stewart et al. (2015). They demonstrated that risk aversion was a property of the choice set used to elicit the measure and not, as is commonly assumed, a stable property of the individual. Our findings here corroborate this conclusion—if you repeat Stewart et al.'s experiment and analysis you will most likely obtain exactly the same results (Alempaki et al. 2019). Thus Stewart et al.'s results are still, as originally claimed, troubling for expected utility theory. But Stewart et al. were wrong to attribute there results to decision by sampling. The results are actually the result of the issue we present here: omitted variable bias differing over data sets.

8.2. But Everything Is Okay If I Am Fitting the Correct Model, Right?

We have shown that the problem is not specific to expected utility theory. The problem actually appears larger for prospect theory. But perhaps there is another model that is the right model? Unfortunately, it is unlikely we can avoid these problems. Given the limits on the number of choices one can ask a participant to make in a choice experiment, there is a limit on the complexity of the model that can be estimated. But there are many known systematic departures from the expected utility model, motivating many descriptive models of risky choice. Starmer (2000) reviewed over 20 models which depart from and extend the expected utility framework, all aiming to capture the systematic departures from expected utility observed in choice experiments presented in the economics literature. Birnbaum (2008) reviewed 11 new systematic departures from prospect theory and expected utility theory which are captured by his transfer-of-attention-exchange model. More recently Bhatia et al. (2016) have counted over 53 descriptive models of risky choice in a systematic review (though many are very closely related), all of which capture different systematic departures from the expected utility model. Any of these many systematic departures from the expected utility model could, potentially, bias parameters estimated for the expected utility model. It seems unlikely—and probably not even desirable—that we will ever develop one model to capture all of these departures. Even if we did, it would probably be too complicated to estimate from sensibly sized choice experiments (though perhaps the hierarchical approach from Bayesian and mixed modeling can assist us by estimating one model across many individuals). But without it, our parameters estimated from the expected utility model will always be biased by the things we have omitted from the model, and the strength of these biases will be different in different choice sets. So this does not end well: either variables are omitted which means the parameter you are measuring is not measuring what you think it is and may not generalize at all well or it means that your model will be too complicated to estimate.

Acknowledgments

This work was supported Economic and Social Research Council grants ES/K002201/1, ES/P008976/1 and ES/K004948/1, and Leverhulme grant RP2012-V-022.

References

Alempaki D, Canic E, Mullett TL, Skylark WJ, Starmer C, Stewart N, Tufano F (2019) Re-examining how utility and weighting functions get their shapes: A quasi-adversarial collaboration providing a new interpretation. *Management Science* 65:4841–4862, URL http://dx.doi.org/10.1287/mnsc.2018.3170.

Allais M (1953) Le comportement de l'homme rationel devant le risque: Critique des postulats et axioms de l'ecole americaine [Rational man's behavior in face of risk: Critique of the American School's postulates and axioms]. *Econometrica* 21:503–546.

Andersson O, Tyran JR, Wengstrom E, Holm HJ (2013) Risk aversion relates to cognitive ability: Fact or fiction?, URL http://dx.doi.org/10.2139/ssrn.2347367.

Bell DE (2007) Utility and risk preferences. Edwards W, Miles RF, von Winterfeldt D, eds., *Advances in decision analysis: From foundations to applications*, 221–231 (New York: Cambridge University Press).

Benjamin DJ, Brown SA, Shapiro JM (2013) Who is 'behavioral'? Cognitive ability and anomalous preferences. *Journal of the European Economics Association* 11:1231–1255, URL http://dx.doi.org/10.1111/jeea.12055.

Bhatia S, Loomes GL, Read D (2016) The established laws of preferential choice behaviour.

Birnbaum MH (2008) New paradoxes of risky decision making. *Psychological Review* 115:453–501, URL http://dx.doi.org/10.1037/0033-295X.115.2.463.

Burks SV, Carpenter JP, Goette L, Rustichini A (2009) Cognitive skills affect economic preferences, strategic behavior, and job attachment. *Proceedings of the National Academy of Sciences of the United States of America* 106:7745–7750, URL http://dx.doi.org/10.1073/pnas.0812360106.

Canic E (2016) *Value is context dependent: On comparison processes and rank order in choice and judgment*. Ph.D. thesis, University of Warwick, Coventry, England.

Croson R, Gneezy U (2009) Gender differences in preferences. *Journal of Economic Literature* 47:448–474, URL http://dx.doi.org/10.1257/jel.47.2.448.

Dohmen T, Falk A, Huffman D, Sunde U (2010) Are risk aversion and impatience related to cognitive ability? *American Economic Review* 100:1238–1260, URL http://dx.doi.org/10.1257/aer.100.3.1238.

Einav L, Finkelstein A, Pascu I, Cullen MR (2012) How general are risk preferences? choices under uncertainty in different domains. *American Economic Review* 102:2606–2638, URL http://dx.doi.org/10.1257/aer.102.6.2606.

Frey R, Pedroni A, Mata R, Rieskamp J, Hertwig R (2017) Risk preference shares the psychometric structure of major psychological traits. *Science Advances* 3:–, URL http://dx.doi.org/10.1126/sciadv.1701381.

Friedman D, Isaac RM, James D, Sunder S (2014) *Risky curves: On the Empirical Failure of Expected Utility* (Abingdon, England: Routledge).

Gäechter S, Johnson EJ, Herrmann A (2007) Individual-level loss aversion in riskless and risky choices, URL https://ssrn.com/abstract=1010597.

Glöckner A, Betsch T (2008) Do people make decisions under risk based on ignorance? an empirical test of the priority heuristic against cumulative prospect theory. *Organizational Behavior and Human Decision Processes* 107:75–95, URL http://dx.doi.org/10.1016/j.obhdp.2008.02.003.

Glöckner A, Pachur T (2012) Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition* 123:21–32, URL http://dx.doi.org/10.1016/j.cognition.2011.12.002.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *American Economic Review* 92:1644–1655, URL http://dx.doi.org/10.1257/000282802762024700.

Holt CA, Laury SK (2005) Risk aversion and incentive effects: New data without order effects. *American Economic Review* 95:902–904, URL http://dx.doi.org/10.1257/0002828054201459.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–291, URL http://dx.doi.org/10.2307/1914185.

Mata R, Josef AK, Samanez-Larkin GR, Hertwig R (2011) Age differences in risky choice: a meta-analysis. *Decision Making Over the Life Span* 1235:18–29, URL http://dx.doi.org/10.1111/j.1749-6632.2011.06200.x.

Mullett TL, Tunney RJ (2013) Value representations by rank order in a distributed network of varying context dependency. *Brain and Cognition* 82:76–83, URL http://dx.doi.org/10.1016/j.bandc.2013.02.010.

Nelson JA (2015) Are women really more risk-averse than men? A re-analysis of the literature using expanded methods. *Journal of Economic Surveys* 29:566–585, URL http://dx.doi.org/10.1111/joes.12069.

Noguchi T, Stewart N (2018) Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological Review* 125:512–544, URL http://dx.doi.org/10.1037/rev0000102.

Pedroni A, Frey R, Bruhin A, Dutilh G, Hertwig R, Rieskamp J (2017) The risk elicitation puzzle. *Nature Human Behaviour* 1:803–809, URL http://dx.doi.org/10.1038/s41562-017-0219-x.

Rabin M (2000) Risk aversion and expected-utility theory: A calibration theorem. *Econometrica* 68:1281–1292.

Rieskamp J (2008) The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34:1446–1465, URL http://dx.doi.org/10.1037/a0013646.

Starmer C (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38:332–382, URL http://dx.doi.org/10.1257/jel.38.2.332.

Stewart N (2009) Decision by sampling: The role of the decision environment in risky choice. *Quarterly Journal of Experimental Psychology* 62:1041–1062, URL http://dx.doi.org/10.1080/17470210902747112.

Stewart N, Chater N, Brown GDA (2006) Decision by sampling. *Cognitive Psychology* 53:1–26, URL http://dx.doi.org/10.1016/j.cogpsych.2005.10.003.

Stewart N, Reimers S, Harris AJL (2015) On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science* 61:687–705, URL http://dx.doi.org/10.1287/mnsc.2013.1853.

Stewart N, Scheibehenne B, Pachur T (2018) Psychological parameters have units: A bug fix for stochastic prospect theory and other decision models.

Stewart N, Simpson K (2008) A decision-by-sampling account of decision under risk. Chater N, Oaksford M, eds., *The probabilistic mind: Prospects for Bayesian cognitive science*, 261–276 (Oxford, England: Oxford University Press), URL http://www.stewart.warwick.ac.uk/publications/.

Stott HP (2006) Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty* 32:101–130, URL http://dx.doi.org/10.1007/s11166-006-8289-6.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5:297–323, URL http://dx.doi.org/10.1007/BF00122574.

von Neumann M, Morgenstern O (1947) *Theory of games and economic behavior* (Princeton, NJ: Princeton University Press), 2nd edition.

Appendix A: Experiment Method

There were three between-participant conditions in the Canic and Stewart experiment. Choices in the Positive and Uniform Conditions followed the recipe in Stewart et al.. For these conditions, the results are a straightforward replication and we do not consider them further.

The third condition was made by mixing together the Positive distribution (of £10, £20, £50, £100, £200, and £500) with the Uniform distribution (of £100, £200, £300, £400 and £500) to give the union distribution containing £10, £20, £50, £100, £200, £300, £400 and £500. When crossed with the probabilities 20%, 40%, 60%, 80%, and 100% this gives 280 choices between non-identical, non-dominating pairs of gambles. We selected a random subset of 140 of these choices, subject to the constraint that each amount, probability, pair of amounts, and pair of probabilities appeared equally frequently. This condition is reported in the meta analysis in Alempaki et al. (2019) as Experiment L3.e.

We decided in advance to recruit 360 participants and collected the data in two simultaneous batches, one from Prolific Academic and the other from Amazon Mechanical Turk. For MTurk, we required participants from the United States, and swapped pound symbols for dollar symbols. 121 participants were assigned to the

mixed condition. We decided in advance to exclude all submissions that were not from unique IP addresses (to address repeat participation by the same individual under multiple accounts) and, to address those not completing the task seriously, the fastest 5% and slowest 5% in each condition and the 5% of people who alternated responses the most (switching between the left and right gambles too much) and the 5% of people who alternated responses too little (repeatedly choosing left too much, or right too much). In examining the total time and alternation distributions we can see that many who were excluded were not that extreme, but a few of the excluded participants were clearly not engaging. Still, we stuck with the in-advance exclusion criteria, which led to the removal of 32 participants from the mixed condition.

In the instructions participants were told to imagine a gamble as a bag of 100 tickets. They were given the concrete example of a choice between a "40% chance of £200" and an "80% chance of £100", which is common to both conditions. For "40% chance of £200" they were told to imagine a bag holding 100 tickets, with 40 winners marked £200 and 60 remaining tickets marked £0. They were told to imagine that the gamble would be resolved by blindly drawing one ticket at random from the bag. They were told that at the end of the experiment we would randomly select four participants and, for each, randomly select one of their choices to play out in this way. We told them that their choices were worth a lot to us, that there were no right answers, and that they could win up to £25 given the experiment exchange rate.

On each trial two buttons were presented, one for each gamble, beneath the question "Which gamble do you prefer?". The text on the button described the gamble with the phrase "p% chance of £x". Participants clicked the gamble they preferred. The entire screen then went blank for 500ms before the next trial.

The experiment took participants between 10 and 20 minutes to complete.