informs
http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Hierarchical Maximum Likelihood Parameter Estimation for Cumulative Prospect Theory: Improving the Reliability of Individual Risk Parameter Estimates

Ryan O. Murphy,[a] **Robert H. W. ten Brincke**[b]

[a] Department of Economics, University of Zürich, 8006 Zürich, Switzerland; [b] ETH Zürich, 8092 Zürich, Switzerland
Contact: ryan.murphy@econ.uzh.ch (ROM); r.h.w.tenbrincke@gmail.com (RHWB)

Received: January 28, 2015
Accepted: June 20, 2016
Published Online in Articles in Advance:

https://doi.org/10.1287/mnsc.2016.2591

Copyright: © 2017 INFORMS

Abstract. An individual's tolerance of risk can be quantified by using decision models with tuned parameters that maximally fit a set of risky choices the individual has made. A goal of this model fitting procedure is to identify parameters that correspond to stable underlying risk preferences. These preferences can be modeled as an individual difference, indicating a particular decision maker's tastes and willingness to accept risk. Using hierarchical statistical methods, we show significant improvements in the reliability of individual risk preference parameter estimates over other common methods for cumulative prospect theory. This hierarchical procedure uses population-level information (in addition to an individual's choices) to break "ties" (or near ties) in the fit quality for sets of possible risk preference parameters. By breaking these statistical ties in a sensible way, researchers can avoid overfitting choice data and thus more resiliently measure individual differences in people's risk preferences.

History: Accepted by Yuval Rottenstreich, judgment and decision making.

Keywords: measuring risk preferences • cumulative prospect theory • decision making under risk • hierarchical parameter estimation • individual differences • risk profile • reliability

1. Introduction

People must often make choices among a number of different options for which the outcomes are not certain. Such choices are referred to as *risky* when the options are well-defined sets of outcomes and each has its respective payoff(s) and probability of fruition (Knight 1921, Edwards 1954, Luce and Raiffa 1957) clearly described. Expected value (EV) maximization stands as a benchmark solution to risky choice problems, but people's behavior does not always conform to this optimization principle. Rather, decision makers (DMs) reveal different preferences for risk, sometimes, for example, forgoing an option with a higher expectation in lieu of an option with lower variance (thus indicating risk aversion; in some cases DMs reveal the opposite preference, too, indicating risk seeking). Behavioral theories of risky decision making (Kahneman and Tversky 1979, 2000; Tversky and Kahneman 1992; Camerer 1995) have been developed to identify and highlight the structure in these choice patterns and provide psychological insights into DMs' revealed preferences. Reliable measures of risk preferences allow researchers to investigate the associations between risky choice behavior and other variables of interest and allow the incorporation of risk preferences in other contexts (e.g., Camerer 2004, Huettel et al. 2006). This paper is about measuring those subjective risk preferences and in particular developing a statistical estimation procedure that can increase the reliability and robustness of those parameter estimates, thus better capturing risk preferences and doing so consistently at the individual level.

1.1. Three Necessary Components for Measuring Risk Preferences

There are three elementary components in measuring risk preferences and these parts serve as the foundation for developing behavioral models of risky choice. These three components are lotteries, models, and statistical estimation/fitting procedures. We explain each of these components below in general terms and then provide detailed examples and a discussion of the elements in the subsequent sections of the paper.

1.1.1. Lotteries. Choices among options reveal DMs' preferences (Samuelson 1938, Varian 2006). Lotteries are used to elicit risky decisions, and these resulting choice data serve as the input for the decision models and parameter estimation procedures. Here we focus on choices from binary lotteries (Stott 2006, Rieskamp 2008, Nilsson et al. 2011; see Table A.1 in Appendix A for an example lottery and Appendix E for all the lotteries). The elicitation of certainty equivalence values is another well-established method for quantifying risk preferences (e.g., see Zeisberger et al. 2012). Binary lotteries are arguably a simpler way for DMs to make risky choices and thus elicit risk preferences. One reason to use binary lotteries is because people appear to have problems with assessing a lottery's certainty equivalent (Lichtenstein and Slovic 1971, Harrison and Rutström 2008), although this simplicity comes at the cost of requiring a DM to make many binary choices to achieve the same degree of estimation fidelity that other methods purport to have (e.g., the three decision risk measure from Tanaka et al. 2010).

The approach we cover here employs a set of static lotteries; this fixed approach has some advantages in its ease of implementation and the comparability of results between subjects. Other approaches have used adaptive lotteries that require sophisticated algorithms to dynamically design risk elicitation items contingent upon what choices DMs have made so far. Cavagnaro et al. (2013) develop an adaptive approach for model discrimination. This is a useful tool when the research goal is to determine if a DM's choices are more consistent with one particular choice model or another. Toubia et al. (2013) have developed a different adaptive approach that aims to maximize the information obtained in each step of a choice experiment involving risk and time preferences. Such adaptive methods appear to be more efficient (ultimately requiring fewer choices from DMs for estimating preferences and thus demanding less time and effort) but are also more challenging to implement, which may act as a barrier to their widespread adoption. Adaptive methods are also predicated on assumptions that preferences are stable and independent of the elicitation process. Other approaches (Payne et al. 1992; Parducci 1995; Stewart et al. 2003, 2006; Lichtenstein and Slovic 2006) posit that often DMs do not in many situations have stable long-term preferences but rather dynamically sample information, make comparative judgments from this particular local context, and then construct preferences in real time. Such a view would also favor using a static set of lotteries over adaptive methods.

1.1.2. Decision Models (Nonexpected Utility Theory).

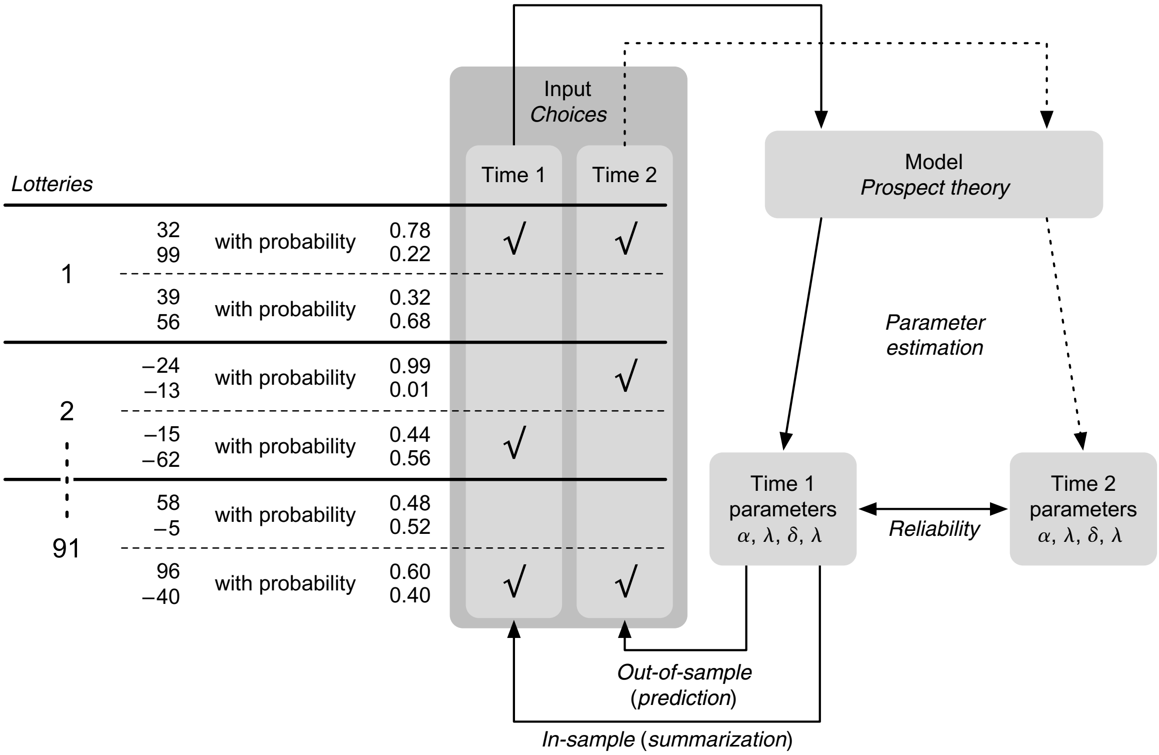
Models may reflect the general structure (e.g., stylized facts) of DMs' aggregate preferences by invoking a latent construct like utility. These models typically also have free parameters that can be tuned to improve the accuracy of how they represent different choice patterns. An individual's choices from the lotteries are fit to a decision model by tuning these parameters, thus identifying individual differences in revealed preferences and summarizing the pattern of choices by using particular combinations of parameter values. These parameters can, for example, not just establish the existence of risk aversion but also quantify the degree of this particular preference. This is useful in characterizing an individual DM and isolating the cognitive mechanisms that underlie behavior (e.g., correlating risk preferences with other variables, including measures of physiological processes such as Huettel et al. 2006, Engelmann and Tamir 2009, Figner and Murphy 2010, or process tracing approaches that track attention as in Schulte-Mecklenbeck et al. 2016) as well as tuning a model to make predictions about what a particular DM will choose when presented with different options. Here we focus on the nonexpected utility model cumulative prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992) as our decision model, using functional specifications explained below. Prospect theory is arguably the most important and influential descriptive model of risky choice to date (Starmer 2000, Wakker 2010, Barberis 2013, Fox et al. 2015), and it has been used extensively in research related to model fitting and risky decision making.

1.1.3. Parameter Estimation Procedure.

Given a set of lotteries, risky choices, and a particular decision model, the best fitting parameter(s) can be identified via a statistical estimation procedure. A perfectly fitting parameterized model would exactly reproduce all of a DM's choices. Maximum likelihood estimation methods have been developed that allow for a more nuanced approach (i.e., not just counting the number of correct predictions) in evaluating the fit of a choice model (see, e.g., Harless and Camerer 1994, Hey and Orme 1994, Regenwetter and Robinson 2016) to behavioral data. Other methods build on this and use mixture models, in which DMs' choices are fit to several models simultaneously rather than only one (see Harrison and Rutström 2009, Conte et al. 2011). Mixture models can also be used to identify the most common preference types by specifying a flexible decision model (Bruhin et al. 2010). This approach can be powerful because clusters of types of DMs can emerge endogenously, yielding a structure to organize the considerable heterogeneity observed in people's risk taking behavior. Regardless of the modeling approach, all of these evaluation methods can include in-sample and out-of-sample tests, the latter of which are especially useful because they can diagnose and mitigate the overfitting of a model to choice data.

The interrelationship between lotteries, a decision model, and an estimation procedure is shown in Figure 1. Lotteries provide stimuli and the resulting choices are the behavioral input for the model; a decision model and an estimation procedure tune parameters to maximize correspondence between the model and the actual choices from the DM facing the lotteries. This process of eliciting choices can be repeated again using the same lotteries and the same experimental subjects. This test-retest design allows for both the

Figure 1. Interrelationship Between Lotteries, a Decision Model, and an Estimation Procedure


Input
Choices
Time 1
Time 2
Model
Prospect theory
Lotteries
1
32 99 with probability 0.78 0.22 √ √
39 56 with probability 0.32 0.68
2
−24 −13 with probability 0.99 0.01 √
−15 −62 with probability 0.44 0.56 √
91
58 −5 with probability 0.48 0.52
96 −40 with probability 0.60 0.40 √ √
Parameter estimation
Time 1 parameters $\alpha, \lambda, \delta, \lambda$
Reliability
Time 2 parameters $\alpha, \lambda, \delta, \lambda$
Out-of-sample (prediction)
In-sample (summarization)

Notes. On the left are choices from binary lotteries that serve as input for the model. Using an estimation method, we fit the model's risk preference parameters (of cumulative prospect theory) to the individual risky choices at time 1. These parameters, to some extent, summarize the choices made in time 1 by using the parameters in the model and reproduce the choices (depending on the quality of the fit). The out-of-sample predictive capacity of the model can be evaluated by comparing the predictions against actual choices at time 2. The reliability of the estimates is evaluated by comparing the time 1 parameter estimates to estimates obtained using the same estimation method on the time 2 data. This experimental design holds the lotteries and choices constant but varies the estimation procedures; moreover, this is done in a test-retest design that allows for the computation and comparison of both in-sample and out-of-sample statistics.

development of in-sample parameter fitting (i.e., statistical "explanation" or summarization) as well as out-of-sample parameter fitting (i.e., prediction). Moreover, the test-retest reliability of the different parameter estimates can be computed as a correlation coefficient because two sets of parameters exist for each DM. This is a powerful experimental design because it can accommodate both in-sample fitting and also prediction and further can diagnose instances of overfitting which can undermine reliability as well as meaningful psychological interpretations (see Pitt and Myung 2002, Lewandowsky and Farrell 2010).

1.2. Overfitting and Bounding Parameters

Multiparameter models' estimation methods may be prone to overfitting and in doing so adjust to noise instead of real risk preferences (Roberts and Pashler 2000). This can sometimes be observed when parameter values emerge that are highly atypical and extreme. A common solution to this problem is to set boundaries and limit the range of parameter values that are potentially estimated. Boundaries prevent extreme parameter values and thus can reduce overfitting, but on the downside they negate the possibility of detecting extreme preferences altogether, even though these preferences may be real but unusual. Boundaries are also defined arbitrarily and may create serious estimation problems due to parameter interdependence. For example, setting a different boundary on one parameter may radically change the estimate of another parameter (e.g., restricting the range of probability distortion may unduly influence estimates of loss aversion) for one particular subject. The interdependence of parameter estimates has been noted (Nilsson et al. 2011, Zeisberger et al. 2012). To circumvent the pitfalls of arbitrary parameter boundaries, we use a hierarchical estimation method based on Farrell and Ludwig (2008) without such boundaries. At its core, this hierarchical method uses estimates of risk preferences of the whole sample to inform estimates of the risk preferences at the individual level. We therefore address to what degree an estimation method combining group-level information with individual-level information can more reliably represent individual risk preferences compared with using either individual or aggregate information exclusively.

1.3. Justification for Using Hierarchical Estimation Methods

The ultimate goal of the hierarchical estimation procedure here is to obtain improvements in the reliability of estimates for individual risk preferences that can be used to make better predictions about risky choice behavior, contrasted to other estimation methods. This is not modeling as a means only to maximize in-sample fit but rather to maximize out-of-sample correspondence; moreover, these methods provide a way to gain insights into what people are actually motivated by as they cogitate about risky options and make choices when confronting irreducible but quantified uncertainty. Ideally, the parameters should not only be about merely summarizing choices but also reflect some psychological mechanisms that underlie risky decision-making behavior.

1.4. Other Applications of Hierarchical Estimation Methods

Other researchers have applied hierarchical estimation methods to risky choice modeling at the individual level. Nilsson et al. (2011) have applied a Bayesian hierarchical parameter estimation model to simple risky choice data. The hierarchical procedure outperformed maximum likelihood estimation in a parameter recovery test. The authors, however, did not test out-of-sample predictions nor the test-retest reliability of their parameter estimates, a gap we address in this paper. Wetzels et al. (2010) applied Bayesian hierarchical parameter estimation in a learning model and found it was more robust with regard to extreme estimates and misunderstandings of the nature of the decision-making task. Scheibehenne and Pachur (2013) applied Bayesian hierarchical parameter estimation to risky choice data using a Transfer of Attention eXchange (TAX) model (Birnbaum and Chavez 1997, Birnbaum 1999) but reported no improvement in parameter stability. In another study, Scheibehenne and Pachur (2015) also find no improvement in parameter stability for prospect theory. The authors offer the explanation that hierarchical estimation at times overcorrects for extreme but correct parameter values and that the benefits of hierarchical estimation may be limited to cases where data are sparse.

1.5. Other Research Examining Parameter Stability

Besides results from the hierarchical approaches outlined above, there exists other research on estimated parameter stability. Glöckner and Pachur (2012) tested the reliability of parameter estimates using a non-hierarchical estimation procedure. The results show that individual risk preferences are generally stable and that the individual parameter values outperform aggregate values in terms of prediction. The authors concluded that the reliability of parameters suffered when extra model parameters were added. This result is contrasted against the work of Fehr-Duda and Epper (2012), who conclude, based on experimental data and a literature review, that those additional parameters (related to the probability weighting function) are necessary to capture individual risk preferences. Zeisberger et al. (2012) also tested individual parameter reliability using a different type of risky choice (certainty equivalent rather than binary choice). They found significant differences in parameter value estimates over time but did not use a hierarchical estimation procedure.

1.6. Major Contributions

In this paper we show concretely how the use of a hierarchical estimation method is able to circumvent many of the problems that occur with maximum likelihood procedures when estimating individual parameters in cumulative prospect theory. The hierarchical method outperforms maximum likelihood estimation with regard to out-of-sample fit and parameter reliability and does not resort to arbitrarily chosen parameter bounds or to the elimination of subjects to accomplish this efficiency gain. As a statistical method it is simpler, more transparent, and computationally less demanding than other approaches. We explore how this method is a more nuanced way of discerning parameters through the use of a hierarchical step, which produces a kind of *plausibility filter* for sets of individual-level parameters. This helps develop an explanation and intuition for how hierarchical methods achieve these superior results. We conclude that the hierarchical maximum likelihood (HML) method is readily applicable "off the shelf," making it a powerful approach for both researchers and practitioners interested in measuring DMs' risk preferences.

1.7. Structure of the Rest of This Paper

In this paper we focus on evaluating different statistical estimation procedures for fitting individual choice data to a risky choice model. To this end, we hold constant the set of lotteries DMs made choices with and the functional form of the risky choice model. We then fit the same set of choice data using two estimation methods and contrast the results to differentiate the quality of the different statistical methods. We also compute the test-retest reliability of individual-level parameters that resulted from different estimation procedures. This broad approach allows us to evaluate different estimation methods and make substantiated conclusions about the fit quality as well as diagnose overfitting. We conclude with recommendations for estimating risk preferences at the individual level.

Table 1. Examples of Binary Lotteries from Each of the Four Types

Lottery	Option A			Option B			Type
1	32 99	with probability	0.78 0.22	39 56	with probability	0.32 0.68	Gain
2	−24 −13	with probability	0.99 0.01	−15 −62	with probability	0.44 0.56	Loss
3	58 −5	with probability	0.48 0.52	96 −40	with probability	0.60 0.40	Mixed
4	−30 60	with probability	0.50 0.50		0 for sure		Mixed-zero

Note. These lotteries were part of the set of 91 lotteries used in the experiment in this paper.

2. Lotteries and Experimental Design

2.1. Participants

One hundred eighty-five participants from the subject pool at the Max Planck Institute for Human Development in Berlin volunteered to participate in two research sessions (referred to hereafter as time 1 and time 2) that were administered approximately two weeks apart. After both sessions were conducted, complete data from both sessions were available for 142 participants; these subjects with complete data sets are retained for subsequent analysis.[1] The experiment was incentive compatible and was conducted without deception. The data used here are the resulting choices from the Schulte-Mecklenbeck et al. (2016) experiment, and we are indebted to these authors for their generosity in sharing their raw data.

2.2. Lotteries

In each session of the experiment, individuals made choices from a set of 91 simple binary lotteries. Each option has two possible outcomes between −100 and 100 that occur with known probabilities that sum to one. There were four types of lotteries for which examples are shown in Table 1: gains only, losses only, mixed lotteries with both gains and losses, and mixed-zero lotteries with one gain and one loss and zero (status quo) as the alternative outcome. The first three types were included to cover the spectrum of risky decisions and the mixed-zero type allows for measuring loss aversion separately from risk aversion (Rabin 2000, Wakker 2005). The same 91 lotteries were used in both test-retest sessions of this research. The set of lotteries was compiled of existing items used by Rieskamp (2008), Gächter et al. (2007), and Holt and Laury (2002). In total, 35 lotteries are gain only, 25 are loss only, 25 are mixed, and 6 are mixed-zero. All of the lotteries are listed in Appendix E.

2.3. Procedure

Participants received extensive instructions regarding the experiment at the beginning of the first session. All participants received EUR 10 as a guaranteed payment for participation in the research and could earn more based on their choices. Participants worked through several examples of risky choices to familiarize themselves with the interface of the MouselabWeb software (Willemsen and Johnson 2011, 2014) that was used to administer the study. Although the same 91 lotteries were used for both experimental sessions, the item order was fully randomized. Additionally, the order of the outcomes (top-bottom) and option order (A or B) and the orientation (A above B, A left of B) were randomized and stored during the first session. The exact opposite spatial representation of all of this was used in the second session to mitigate potential order or presentation effects.

Incentive compatibility was implemented by randomly selecting one lottery at the end of each experimental session, playing it out with the stated probabilities, and paying the participant according to her choice and the realized outcome on the selected item. An exchange rate of 10:1 was used between experimental values and payments for choices. Thus, all participants earned their fixed payment plus one tenth of the outcome of one randomly selected lottery for each completed experimental session. Subjects knew all of this information. Participants earned in total about EUR 30 (approximately USD 40) on average for participating in both sessions.

3. Model Specification

We use cumulative prospect theory (Tversky and Kahneman 1992) to model risk preferences. A two-outcome lottery L is valued in utility $u(\cdot)$ as the sum of its components in which the monetary reward x_i is weighted by a value function $v(\cdot)$ and the associated probability p_i, which is transformed by a probability weighting function $w(\cdot)$. This is shown in the following equation:[2]

$$u(L) = \begin{cases} v(x_1)w(p_1) + v(x_2)(1 - w(p_1)) \\ \quad \text{if both positive/both negative,} \\ \quad |x_1| > |x_2|; \\ v(x_1)w(p_1) + v(x_2)w(p_2) \quad \text{if mixed.} \end{cases} \tag{1}$$

Cumulative prospect theory has many possible mathematical specifications. These different functional specifications have been tested and reported in the literature (see, e.g., Stott 2006) and the functional forms and parameterizations outlined below are justifiable given the preponderance of previous findings.

3.1. Value Function

In general, power functions have been shown to fit behavioral choice data better than many other functional forms at the individual level (Stott 2006). Stevens (1957) also cites experimental evidence showing the merit of power functions in general for modeling psychological processes. Here we use a power value function as displayed in the following equation:

$$v(x) = \begin{cases} x^{\alpha} & \text{if } x \geq 0, \quad \alpha > 0; \\ -\lambda(-x)^{\alpha} & \text{if } x < 0, \quad \alpha > 0; \lambda > 0. \end{cases} \tag{2}$$

The α parameter controls the curvature of the value function. If the value of α is below one, there are diminishing marginal returns in the domain of gains. If the value of α is one, the function is linear and consistent with risk neutrality (i.e., EV maximizing) in decision making. For α values above one, there are increasing marginal returns for positive values of x. This pattern reverses in the domain of losses (values of x less than zero), which is referred to as the reflection effect.

For this paper we use the same value-function parameter α for both gains and losses, and our reasoning is as follows. First, this is parsimonious. Second, when using a power function with different parameters for gains and losses, loss aversion cannot be defined anymore as the magnitude of the kink at the reference point (Köbberling and Wakker 2005) but would instead need to be defined contingently over the whole curve.[3] This complicates interpretation and undermines clear separability between utility curvature and loss aversion; further, it may cause modeling problems for mixed lotteries that include an outcome of zero. Problems of induced correlation between the loss aversion parameter and the curvature of the value function in the negative domain have furthermore been reported when using $\alpha_{gain} \neq \alpha_{loss}$ in a power utility model (Nilsson et al. 2011).

3.2. Probability Weighting Function

To capture subjective probability distortion, we use Prelec's functional specification (Prelec 1998; see Figure B.1 in Appendix B). Prelec's two-parameter probability weighting function can accommodate a wide range of curves and it has been shown to fit individual data well (Gonzalez and Wu 1999, Fehr-Duda and Epper 2012). Its specification[4] can be found in the following equation:

$$w(p) = \exp\left(-\delta(-\ln(p))^{\gamma}\right), \quad \delta > 0; \gamma > 0. \tag{3}$$

The γ parameter controls the curvature of the probability weighting function. The psychological interpretation of the curvature of the function is a diminishing sensitivity away from the end points: both zero and one serve as necessary boundaries and the further from these edges, the less sensitive individuals are to changes in probability (Tversky and Kahneman 1992). The δ parameter controls the general elevation of the probability weighting function. It is an index of how attractive lotteries are in general (Gonzalez and Wu 1999) and it corresponds to how optimistic or pessimistic an individual DM is.

The use of Prelec's two-parameter weighting function rather than the original specification used in cumulative prospect theory (Tversky and Kahneman 1992) requires additional explanation. In the original specification, the point of intersection with the diagonal changes simultaneously with the shape of the weighting function, whereas in Prelec's specification, the weighting line always intersects at the same point if δ is kept constant.[5] However, changing the point of intersection and curvature simultaneously induces a negative correlation with the value-function parameter α because both parameters capture similar characteristics; moreover, this specification does not allow for a wide variety of individual preferences (see Fehr-Duda and Epper 2012). Furthermore, the original specification of the weighting function is nonmonotonic for $\gamma < 0.279$ (Ingersoll 2008). Although that low value is generally not in the range of reported aggregate parameters (Camerer and Ho 1994), this nonmonotonicity may become relevant and problematic when estimating individual preferences, where there is considerably more heterogeneity and noise in the choice data.

4. Estimation Methods

Individual risk preferences are captured by obtaining a set of parameters that best fit the observed choices implemented via a particular choice model. In this section we explain the workings of standard maximum likelihood estimation (MLE) and of hierarchical maximum likelihood estimation (HML).

4.1. Maximum Likelihood Estimation

A narrow interpretation of cumulative prospect theory dictates that even a trivial difference in utility leads the DM to *always* choose the option with the highest utility. However, even in early experiments with repeated lotteries, Mosteller and Nogee (1951) found that this is not the case with real DMs. Choices were instead partially stochastic, with the probability of a DM choosing the generally more favored option increasing as the utility difference between the options increased. This idea of *random utility maximization* has been developed by Luce (1959) and others (e.g., Luce and Suppes 1965, McFadden 1980, Harless and Camerer 1994, Hey and

Orme 1994, Loomes and Sugden 1995). In this tradition, we use a generalized logistic function that specifies the probability of picking one option, depending on each option's utility (i.e., softmax). This formulation is displayed in Equation (4). There are other stochastic choice functions that have been used in the literature (see Stott 2006, Table 4). A logistic function has a long history, being used to fit choice data from Mosteller and Nogee (1951), and has been shown generally to perform well with behavioral data (Stott 2006):

$$p(A \succ B) = \frac{1}{1 + e^{\varphi(u(B)-u(A))}}, \quad \varphi \geq 0. \tag{4}$$

The parameter φ is an index of the sensitivity to differences in utility. Lower values for φ diminish the importance of the difference in utility. It is important to note that this parameter operates on the absolute difference in utility assigned to both options. Two individuals with different risk attitudes, but equal sensitivity, will almost certainly have different φ parameters simply because the differences in the options' utilities are also determined by their risk attitude. Although the parameter is useful to help fit an individual's choice data, one cannot compare the values of φ across individuals unless *both* the lotteries *and* the individuals' risk attitudes are held constant. The interpretation of the sensitivity parameter is therefore ambiguous and it should be here considered simply as an aid in the estimation method.

In maximum likelihood estimation the goal function is to maximize the likelihood of observing the outcome, which consists of the observed choices, given a set of parameters. The likelihood is expressed using the choice function above, yielding a stochastic specification. We use the notation $\mathscr{M} = \{\alpha, \lambda, \delta, \gamma, \varphi\}$, where y_i denotes the choice for the ith lottery:

$$\hat{\mathscr{M}}_i = \arg\max_{\mathscr{M}} \prod_{i=1}^{N} c(y_i \mid \mathscr{M}). \tag{5}$$

By $c(\cdot)$, we denote the choice itself, where $p(A \succ B)$ is given by the logistic choice function in Equation (4):

$$c(y_i \mid \mathscr{M}) = \begin{cases} p(A \succ B) & \text{if A is chosen } (y_i \text{ is } 0), \\ 1 - p(A \succ B) & \text{if B is chosen } (y_i \text{ is } 1). \end{cases} \tag{6}$$

MLE takes the utility difference between the two options into account (see Equation (4)) and therefore does not only fit the number of choices correctly predicted but also is sensitive to the magnitude of the difference between the options' utilities. Ironically, because it is the *utility difference* that drives the fit (and not the frequency of correct predictions), the resulting best-fitting parameters may explain *fewer* choices than can be obtained by finding parameters that maximize the fraction of explained choices. This is because MLE tolerates several smaller prediction mistakes instead of fewer, but very large, mistakes. The maximum likelihood quality of the fit drives the parameter estimates, and this improves out-of-sample performance over methods that aim to maximize the fraction of explained choices in-sample.

On the downside, MLE is not robust to aberrant choices. If by confusion or accident a DM chooses an atypical option for a lottery that is greatly out of line with her other choices, the resulting parameters may be disproportionately affected by this single anomalous choice. Although MLE does take into account the quality of the fit with respect to utility discrepancies, the fitting procedure has the simple goal of finding the parameter set that generates the best overall fit (even if the differences in fit among parameter sets are negligible). This approach ignores the reality that there may be different parameter sets that fit choice sets virtually the same. Consider estimations for subject A in Figure 4 (see Section 6.3), which shows the resulting distribution of parameter estimates when one single choice is changed and then the MLE parameters are reestimated. Especially for loss aversion and the elevation of the probability weighting function, these sensitivity analysis results show that a different response on a *single* choice can make a large difference among the resulting best fitting parameter values.

The MLE method solves parameter tie-breaking problems but does so by ascribing a great deal of importance to potentially trivial differences in fit quality. Radically different combinations of parameters may be identified by the procedure as nearly equivalently good, and the winning parameter set may emerge but only differ from the other "almost-as-good" sets by the slightest of margins in fit quality. This process of selecting the winning parameter set from among the space of possible combinations ignores how plausible the sets of parameters may be in capturing the DM's actual preferences. Moreover, the ill-behaved structure of the parameter space makes it a challenge to find reliable results because different solutions can be nearly identical in fit quality but be located in far distant regimes of the ill-behaved parameter space. A minor change in the lotteries, or just one different choice, may bounce the resulting parameter estimations around substantially. Furthermore, the MLE procedure does not consider the psychological interpretability or plausibility of the parameters and therefore may yield a parameter set that is grossly atypical, doing so with only very weak evidence to justify the extreme conclusion.

Results for MLE using risky binary choice problems may lead to particularly lumpy likelihood surfaces. To prevent the evaluation algorithm from becoming stuck in local minima, a large number of starting solutions was used. For each subject, the likelihood of a grid of starting values was evaluated. The resolution in the range of commonly observed values was chosen to be higher than those outside of it. The 25 solutions for which the likelihood was highest for that particular subject were used as starting points for the Nelder-Mead algorithm. Additionally, a wider grid of 216 starting solutions that covered a large range of parameter values was used for all subjects. Thus, a total of 241 starting points was used to estimate MLE parameters.[6]

4.2. Bounded MLE

A potential weakness of MLE is that only the likelihood drives the parameter estimations; thus, the method sometimes returns parameter estimates that are considered outliers. To circumvent this while avoiding the complete removal of subjects from the analysis, we applied an established practice of using (somewhat arbitrarily chosen) bounds to the MLE estimation method. These bounds are $0.2 \le \alpha \le 2$, $0.2 \le \lambda \le 5$, $0.2 \le \delta \le 3$, and $0.2 \le \gamma \le 3$. They have been chosen such that extreme values are excluded while still allowing for a wide range of risk preferences. This is important because using too narrow of bounds can adversely affect the validity of the estimation procedure.

4.3. Hierarchical Maximum Likelihood Estimation

The HML estimation procedure developed here is based on the work of Farrell and Ludwig (2008). It is a two-step procedure that is explained below.

Step 1. First, the likelihood of occurrence for each of the four model parameters in the population as a whole is estimated. These likelihoods of occurrence are captured using probability density distributions and reflect how likely a particular parameter value is in the population, given everyone's choices. These density distributions are found by solving the integral in Equation (7). For each of the four cumulative prospect theory parameters, we use a log-normal (denoted $\mathrm{LN}(\cdot)$) density distribution since this distribution has only positive values, is positively skewed, has only two distribution parameters, and is not too computationally demanding. Because the sensitivity parameter φ is not independent of the other parameters and because it has no clear psychological interpretation, we do not estimate a distribution of occurrence for φ but instead take the value we find for the aggregate data with classical maximum likelihood estimation for this step.[7] We use the notation $\mathscr{M} = \{\alpha, \lambda, \delta, \gamma, \varphi\}$, $\mathscr{S}_\alpha = \{\mu_\alpha, \sigma_\alpha\}$, $\mathscr{S}_\lambda = \{\mu_\lambda, \sigma_\lambda\}$, $\mathscr{S}_\delta = \{\mu_\delta, \sigma_\delta\}$, $\mathscr{S}_\gamma = \{\mu_\gamma, \sigma_\gamma\}$, and $\mathscr{S} = \{\mathscr{S}_\alpha, \mathscr{S}_\lambda, \mathscr{S}_\delta, \mathscr{S}_\gamma\}$; S and N denote the number of participants and the number of lotteries, respectively:

$$\hat{\mathscr{S}} = \arg\max_{\mathscr{S}} \prod_{s=1}^{S} \iiiint \left[\prod_{i=1}^{N} c(y_{s,i} \mid \mathscr{M})\right] \cdot \mathrm{LN}(\alpha \mid \mathscr{S}_\alpha)\mathrm{LN}(\lambda \mid \mathscr{S}_\lambda)\mathrm{LN}(\delta \mid \mathscr{S}_\delta) \cdot \mathrm{LN}(\gamma \mid \mathscr{S}_\gamma)\, d\alpha\, d\lambda\, d\delta\, d\gamma. \tag{7}$$

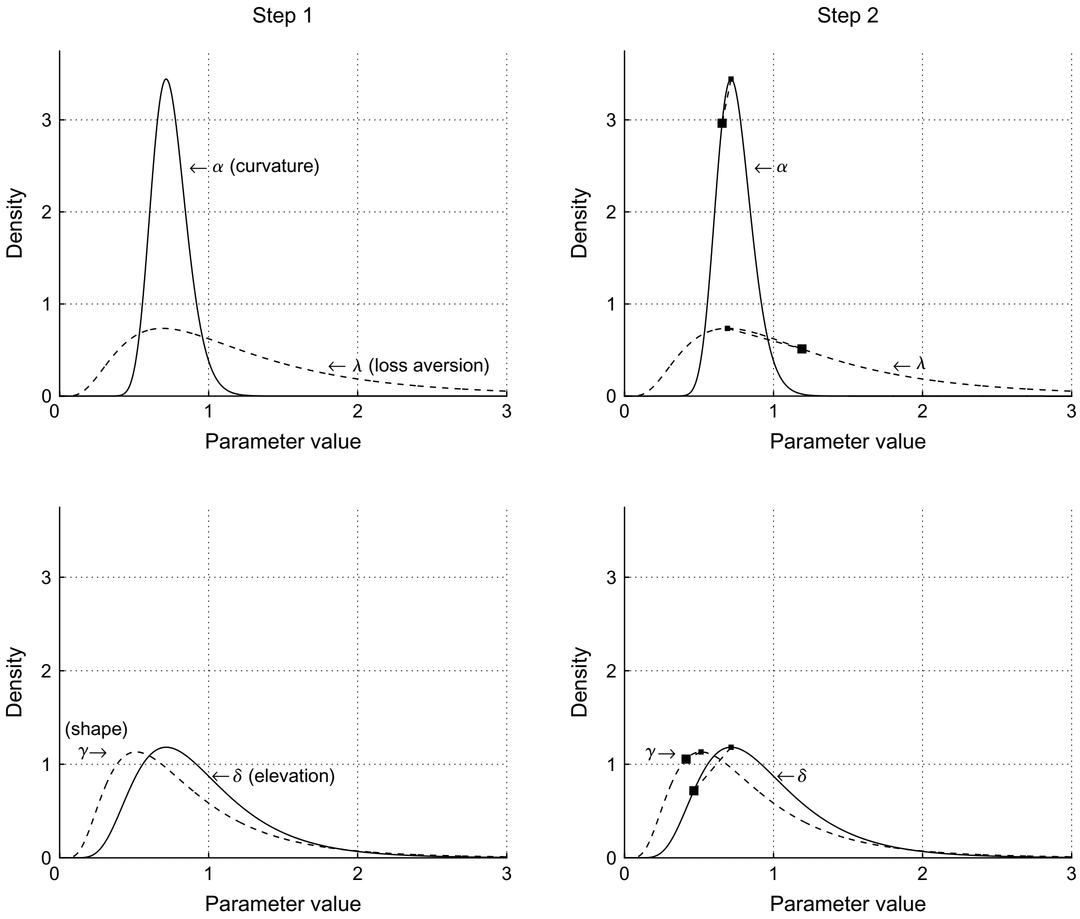
This first step can be explained using a simplified example that is a discrete version of Equation (7). For each of the four risk preference parameters let us pick a set of values that covers a sufficiently large interval, for example $\{0.5, 0.6, 0.7, \ldots, 3.0\}$. We then take all possible parameter combinations (i.e., the Cartesian product) of these four sets, i.e., $\alpha = 0.5$, $\lambda = 0.5$, $\delta = 0.5$, $\gamma = 0.5$, then $\alpha = 0.5$, $\lambda = 0.5$, $\delta = 0.5$, $\gamma = 0.6$, and so forth. Each of these combinations has a certain likelihood of explaining a subject's observed choices, displayed within the block parentheses in Equation (7). However, not all of the parameter values in these combinations are equally supported by the data. It could be, for example, that the combination above that contains $\gamma = 0.5$ is 10 times more likely to explain the observed choices than $\gamma = 1.0$. How likely each parameter value is in relation to another value is precisely what we intend to quantify. The fact that one set of parameters is associated with a higher likelihood of explaining observed choices (and therefore is more strongly supported by the data) can be reflected by changing the four log-normal density functions for each of the model parameters. We multiply the likelihood that a set of values explains the observed choices by the weight put on this set of parameters through density functions; Equation (7) achieves this. The density functions, and the product thereof, denoted by $\mathrm{LN}(\cdot)$ within the integrals, are therefore ways to assess the relative plausibility of particular parameter combinations. The maximization is done over all participants simultaneously under the condition that it is a probability density distribution. Applying this step to our data leads to the density distributions displayed in Figure 2.

Step 2. In this step, we estimate individual parameters as carried out in standard maximum likelihood estimation but now weigh these parameters by the likelihood of their co-occurrence as given by the density functions obtained in Step 1. Those distributions and both steps are illustrated in Figure 2. This second step of the procedure is described in Equation (7). With $\mathscr{M}_i$ we denote $\mathscr{M}$ as above for subject i. The $\hat{\mathscr{S}}$ values used in Equation (8) are from the output of Equation (7). The resulting parameters are driven not only by individual decision data but also by how likely it is that such a parameter combination for an individual occurs in the population:

$$\hat{\mathscr{M}}_i = \arg\max_{\mathscr{M}_i} \left[\prod_{i=1}^{N} c(y_i \mid \mathscr{M}_i)\right] \mathrm{LN}(\alpha \mid \hat{\mathscr{S}}_\alpha)\mathrm{LN}(\lambda \mid \hat{\mathscr{S}}_\lambda) \cdot \mathrm{LN}(\delta \mid \hat{\mathscr{S}}_\delta)\mathrm{LN}(\gamma \mid \hat{\mathscr{S}}_\gamma). \tag{8}$$

This HML procedure is motivated by the principle that *extreme conclusions require extreme evidence*. In the first step of the HML procedure, one simultaneously extracts density distributions of all parameter values using all choice data from the population of DMs. In

Figure 2. A Simplified Example of the Two Steps of the Hierarchical Maximum Likelihood Estimation Procedure



Notes. In the first step, we fit the data to four probability density distributions. In the second step, one obtains the individual estimates. In the absence of individual choice data, the small square has the highest likelihood for any individual, which is at the modal value of each population-level-parameter distribution. The individual-level parameters are then considered in the context of their likelihood of occurrence given Step 1 results, and the emerging best-fitting individual parameter is depicted by the big squares. It is worth noting that this visual representation is a simplification, as the HML procedure is multidimensional in practice (see Equations (7) and (8)).

the second step, the likelihood of a particular parameter value is determined by how well it fits the observed choices from an individual, and at the same time, it is weighted by the likelihood of observing *that particular parameter set* in the population distribution (defined by the density function obtained in Step 1). The parameter set with the most likely combination of both of those steps is selected. The weaker the evidence is, the more extreme parameter estimates are “filtered,” resulting in more plausible parameter sets nearer the center of the density distribution being favored. HML therefore counters maximum likelihood estimation’s tendency to fit parameters to extreme choices by requiring stronger evidence to establish extreme parameter estimates. Thus, if a DM makes very consistent choices, the population parameter densities will have very little effect on the estimates. Conversely, if a DM has greater inconsistency in his choices, the population parameters will have more influence on the individual estimates. In the extreme, if a DM has wildly inconsistent choices, the best parameter estimates for the DM will simply be those for the population average.

The estimation method has been implemented separately in both MATLAB and C. Parameter estimates in the second step were found using the simplex algorithm by Nelder and Mead (1965) and initialized with multiple starting points to avoid local minima. Negative log-likelihood was used for computational stability when possible. For the first step of the hierarchical method, negative log-likelihood was not used for the whole equation because of the presence of the integral, so negative log-likelihood was only used for the multiplication across participants. Solving the volume integral is computationally intense because of the number of dimensions and the low probability values therein. Quadrature methods were found to be prohibitively slow. Because the integral-maximizing parameters are what is of central importance, and not the precise value of the integral, we used Monte Carlo integration with 2,000 (uniform random) values for each dimension as a proxy. We repeated this 50 times and used the average of these estimates as the final distribution parameters.[8] Another possible benefit of HML is that the Step 1 distributions appear to be very stable. With different populations it may be possible to simply take the distributions found in previous studies, whereas other populations may require their own aggregate estimations.[9]

5. Parameter Recovery Results

In parameter recovery procedures, simulated choices are generated using a model and known input parameters. These resulting choices are subsequently used to estimate parameters and the correspondence between the generating parameters and recovered parameters provides insight into the extent to which the estimation procedure's resulting output corresponds to the latent parameters in general. It is a useful way to ascertain whether an estimation method is able to identify parameters as intended.

The results of two parameter recovery procedures are reported here. In the first recovery procedure, 142 simulated DMs were assigned the same parameters, those corresponding to the empirical aggregate parameter values. The noise parameter was drawn randomly from a uniform distribution between 0.2 and 0.4. For each lottery, a choice probability was generated using Equation (4), and a choice was made randomly based on this probability. Parameters were then estimated based on these choices using three estimation methods (MLE, bounded MLE, and HML). This procedure was repeated 100 times. The mean squared error between the actual parameters and the recovered parameters was lowest for HML (by a factor of at least seven to eight for any other parameter). The mean squared error was highest for MLE. Median estimates for prospect theory's parameters were generally slightly lower than the actual parameter values, particularly for α (0.72 for MLE and bounded MLE and 0.69 for HML; 0.73 is the actual value) and λ (1.09 for MLE and bounded MLE, 1.05 for HML; 1.11 is the actual value).[10] In the second recovery procedure, 142 simulated DMs were assigned parameters that were generated randomly from a uniform distribution between the limits of the bounded MLE. This generates a wide range of parameter values, and the procedure is designed to test the flexibility of the estimation methods, particularly for HML. Procedure two was otherwise the same as procedure one, including its uniform sampling range of 0.2 to 0.4 for the noise parameter. Mean squared errors between actual and estimated parameters were lowest for bounded MLE and HML (though higher for MLE). Test-retest correlations between actual and estimated parameters were highest for bounded MLE and HML (0.89, 0.81, 0.83, and 0.85 for α, λ, δ, and γ, respectively). No significant differences emerge between bounded MLE and HML. This shows that despite values being drawn from a uniform distribution, HML is flexible and robust enough to deal with it. These parameter recovery results are consistent with the empirical results, which we turn to next.

6. Empirical Results

6.1. Aggregate Results

The aggregate data can be analyzed to establish stylized facts and general behavioral tendencies. In this analysis, all 12,922 choices (142 participants × 91 lotteries) are modeled together and one set of risk preference parameters is estimated. This preference set is estimated using estimation methods discussed before (MLE, HML). Five parameters (the fifth being the softmax sensitivity parameter φ) are estimated for each method. Note that this analysis, one in which all choices are pooled together, is not suitable for identifying individual risk preferences but rather yields aggregate and stylized results of DMs in general.

The results of the estimation procedures for (α, λ, δ, γ, φ) are $(0.73, 1.11, 0.88, 0.65, 0.30)$ for time 1 and $(0.73, 1.18, 0.84, 0.68, 0.29)$ for time 2. Aggregate parameters are generally quite stable. On the aggregate, we observe behavior consistent with curvature parameter values implying diminishing sensitivity to marginal returns; α values less than one are estimated for both methods, namely, 0.73. Choice behavior is also consistent with some loss aversion ($\lambda > 1$), consistent with the notion that "losses loom larger than gains" (Kahneman and Tversky 1979, p. 279). The highest loss aversion value we find for either of the estimation methods is 1.11 (and 1.18 at time 2), which is much lower than the value of 2.25 reported in the seminal cumulative prospect theory paper by Tversky and Kahneman (1992). Part of the explanation for a lower value (indicating less loss aversion) might be that it is taboo to take money from participants in the laboratory and thus "losses" were only reductions in a participant's show-up payment. The probability weighting function has the characteristic inverse S-shape that intersects the diagonal at around $p = 0.5$. Similar values for all of the estimation methods are found in the second experimental session. The largest difference between the experimental test-retest sessions is seen in probability weighting, where two parameters control a single function and thus may be jointly more sensitive than other parameters.

With the choice function and obtained value of φ, we can convert the parameter estimates of the first experimental session into a prediction of the probability of picking option B over option A for each lottery in the second experimental session (see Equation (4)). This is then compared to the fraction of subjects actually picking option B over option A. On the aggregate, cumulative prospect theory's predictions appear to perform well. The correlation between the predicted and observed probability of picking A over B is 0.93. Details are shown in Appendix D.

6.2. Individual Risk Preference Estimates

When benchmarking the explained fraction of choices and the fit of individual parameters compared to their estimated aggregate counterparts (complete pooling of choices, where each individual is assigned the aggregate parameters), we find that individual parameters provide clear benefits because both the explained fraction of choices and the fit are substantially improved by using individual-level parameters. After having established that individual parameters have clear value in terms of model fit, a more direct comparison between MLE and HML is warranted. Particularly for HML, it is important to determine to what degree its parameter estimates sacrifice in-sample performance, which is the performance of the individual parameters of the first session in terms of how well the parameters summarize the actual choice data. Individual risk preference parameters were estimated for each subject using both estimation methods. Fits are expressed in deviance[11] (lower is better because this value is twice the negative log-likelihood) and are directly comparable because the hierarchical component is excluded from the deviance values for evaluation purposes. The mean and standard deviation of the subject's fits are displayed in Table 2. As to be expected, the log-likelihood is lowest for MLE because that method attempts to find the set of parameters with the best possible fit in-sample. However, HML is very close behind, and more importantly it has substantial advantages in terms of reliability, which we explore in the next section.

We also determined the explained fraction of choices for each individual. The means and standard deviations are displayed in Table 2. Both methods explain 76% of choices. For some participants the HML estimates explain a larger proportion of choices at time 1. Although this seems counterintuitive, it is the fit (likelihood) that drives the MLE parameter estimates and not the fraction of explained choices. For most participants, the HML parameter estimates are different from the MLE estimates, but the different estimates do not result in a different percentage of explained choices.

Table 2. Mean and Standard Deviation (in Parentheses) for the Three Estimation Methods

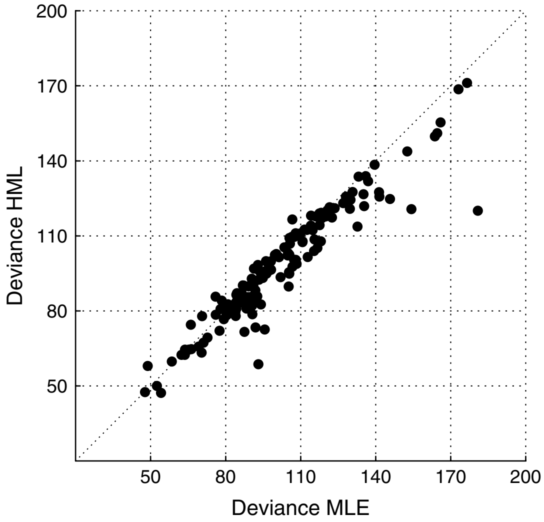
	Explained (time 1)	Predicted (time 2)
Fraction		
MLE	0.76 (0.09)	0.73 (0.09)
MLE (bounded)	0.76 (0.09)	0.73 (0.09)
HML	0.76 (0.08)	0.73 (0.09)
Fit		
MLE	83.96 (21.13)	103.11 (27.42)
MLE (bounded)	85.00 (20.92)	101.77 (26.45)
HML	86.64 (21.66)	99.31 (24.99)

Notes. Explained refers to the time 1 statistics given that time 1 estimates are used, *predicted* refers to time 2 statistics using parameters from time 1. The explained fraction of choices is approximately the same. The fit at time 2 is significantly better for the HML method and indicated by using signed rank tests.

So far we have established that the HML estimates do not hurt the explained fraction of choices accounted for, even if it has a somewhat lower fit in-sample. The next and arguably more important question is whether HML is worth the effort in terms of prediction (i.e., does it perform better out-of-sample). The estimates of individual parameters of the first session were used to predict choices in the second session. In Table 2 we have listed simple statistics for the fit (again comparable by excluding the hierarchical component) and the explained fractions of all three estimation methods for the second session. Both methods explain approximately the same fraction of choices that people make. It is interesting to compare this with the baseline prediction that participants make the same choices in both sessions. That approach correctly predicts 71% on average. Judging by the fraction of lotteries explained, it is not immediately obvious which method is better. A more useful and sensitive measure to compare the estimation methods is the goodness-of-fit statistic. Figure 3 compares the fit for MLE and HML parameters. When comparing HML to MLE, we find that HML's out-of-sample fit is better on average and its deviance is lower for 92 out of 142 participants. Using a paired sign rank test for medians, we find that the fit for HML is significantly lower (better) than for MLE and bounded MLE ($p < 0.001$). The results are perhaps clearer when examining the fit in the figure, which shows many points around the diagonal and the majority below it (lower deviance is better). This shows that HML sacrificed only a small part of its (in-sample) fit at time 1, for an improved (out-of-sample) fit at time 2.

Examination of the differences in parameter estimates for those participants for which HML outperforms MLE reveals higher MLE mean parameter estimates for the individual probability weighting function parameters. The use of bounded MLE, where this issue is reduced, did not change the results because HML still outperformed MLE with boundaries. In many instances, HML offers an improvement over MLE for seemingly plausible MLE estimates, so HML's effect is not limited to outliers or only extreme parameter values. Interestingly, some of the cases in which MLE outperforms HML are those in which extreme low values emerge that are consistent with choices in the second experimental session, such as loss seeking (less weight on losses relative to gains) or weighting functions that are nearly binary step functions. This lack of sensitivity highlights one of the potential downsides of HML because it may pull its estimates toward

Figure 3. The Goodness-of-Fit at Time 2 When Using (Bounded) MLE Parameters from Time 1 (x axis) and HML Parameters from Time 1 (y axis)



Notes. Each point represents one DM. MLE performs better for a DM whose point appears in the upper left triangle; HML performs better when it appears in the lower right triangle.

the peak of the aggregate distribution too much in some cases.

6.3. Parameter Reliability

A key issue in measuring preferences is the reliability of the parameter estimates. This viewpoint is consistent with considering risk preferences as a trait, a stable characteristic of an individual DM, and amenable to psychological interpretation. Reliability can be assessed by applying the different estimation procedures independently on the same data from the two experimental sessions. Large changes in parameter values between time 1 and 2 are a problem and indicate that we are mostly measuring/modeling noise instead of real individual risk preferences. At the individual level we can examine the test-retest correlations of the parameters from the estimation methods, displayed in Table 3. MLE provides rather unreliable estimates. The test-retest reliability for MLE can be improved by applying admittedly somewhat arbitrary bounds. The highest test-retest correlations are obtained using HML, which yields the most reliable parameter estimates for all parameters. Additionally, bootstrapping on the differences in correlation between HML and the other two methods reveals HML's estimates are statistically more reliable for the α, δ, and γ parameters (with 95% intervals).

Table 3. Test-Retest Correlation Coefficients (r Values) for Parameter Values Obtained Using the Listed Estimation Procedures for Both Sessions Independently

	r_α	r_λ	r_δ	r_γ
Correlations				
MLE	0.25	0.37	0.12	0.19
MLE (bounded)	0.29	0.47	0.12	0.43
HML	0.46	0.56	0.49	0.67

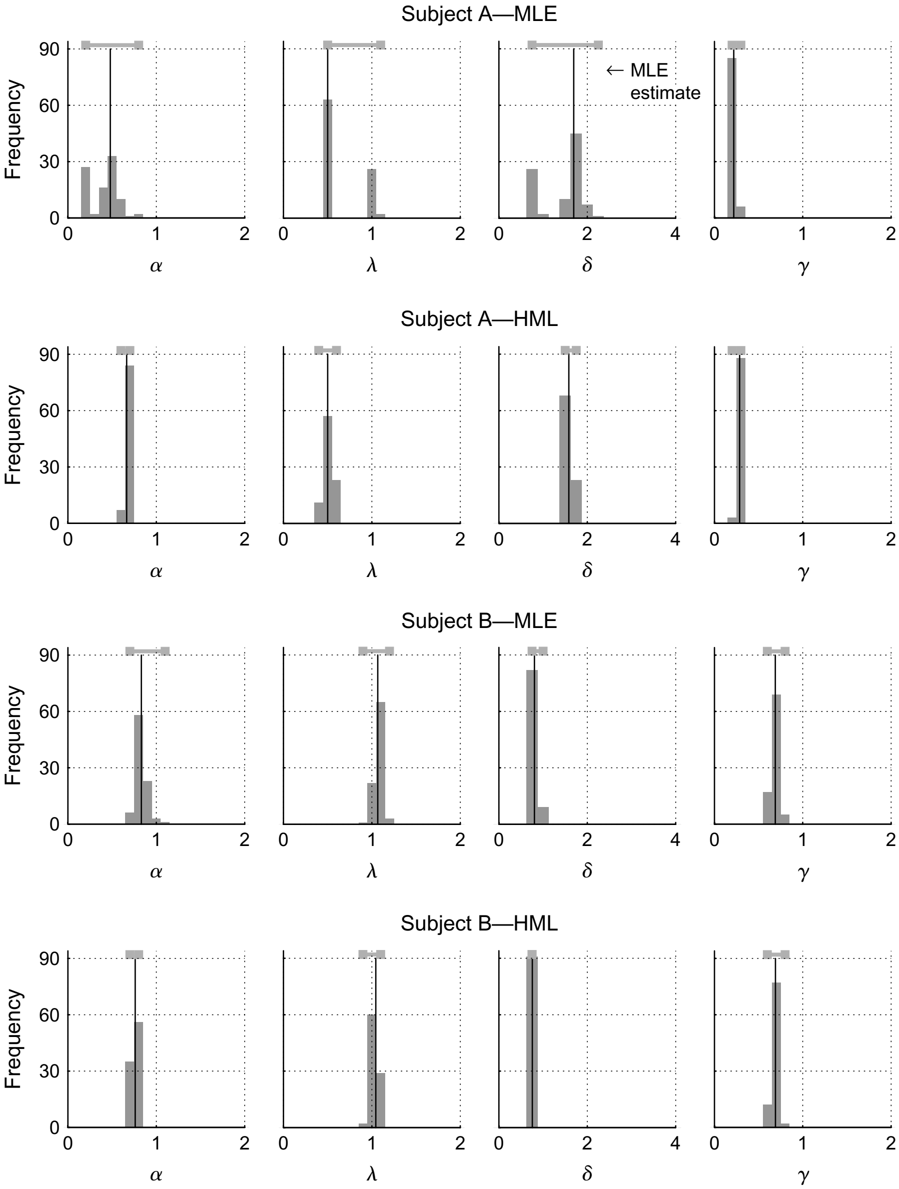
Notes. The use of parameter bounds $0.2 \leq \alpha \leq 2$, $0.5 \leq \lambda \leq 5$, $0.2 \leq \delta \leq 3$, $0.2 \leq \gamma \leq 3$ improves the test-retest correlation for MLE. The highest test-retest correlations are obtained with HML. Bootstrapping was used to test the significance of the increases in correlation that result from using HML. Significant improvements occur for α, γ, and δ over both MLE methods.

Another form of parameter reliability is the change in parameter estimates resulting from small changes in choice input. One way of establishing this form of reliability is by temporarily changing a choice on one of a subject's lotteries, after which the parameters are reestimated. This is repeated for each of the 91 lotteries separately and one at a time. HML's estimates appear to be fairly robust to this procedure, whereas MLE's estimates are less so. In Figure 4 the results of reestimation after perturbing one choice is plotted for two representative subjects using both MLE and HML. A single different choice can yield large changes in MLE parameter estimates and this contributes to unreliable parameter estimates. This fragility is an undesirable property because minor inattention, confusion, or a mistake from a DM would lead to very different conclusions about that person's risk preferences. The robustness of HML estimates is a general finding among subjects in our sample. Moreover, the delicateness of the MLE estimates quickly becomes more problematic as the number of aberrant choices increases by only a few more. In these noisier instances, HML estimates continue to be more resilient.

7. Discussion

This paper illustrates the merits of a hierarchical maximum likelihood (HML) procedure in estimating individual risk parameters. The HML method does reduce

Figure 4. Parameter Outcomes of MLE Parameter Estimates for Two Typical Subjects If There Is a Change to a Single Choice for Each of the 91 Items



Notes. The black line corresponds to the best MLE estimate of the original choices. The light gray horizontal line shows the range of estimations that could emerge for each parameter given one different choice. The multimodal pattern we observe for loss aversion and the weighting function for subject A is not a desirable feature. A momentary lapse of attention or a minor mistake from a DM may result in very different estimated parameter combinations. Compared to MLE, HML produces more reliable estimates by being less sensitive to one aberrant choice.

the in-sample performance, albeit trivially, while at the same time significantly improving the out-of-sample fit compared to MLE. Of central importance, HML estimation yields parameter estimates with higher reliability than MLE. In terms of the bias-variance trade-off often discussed in statistical model evaluation, HML dominates HML by having equivalent in-sample performance, better out-of-sample performance, and better parameter reliability, all while using the exact same data and choice model. If improvements had occurred in either out-of-sample fit or parameter reliability, but not both simultaneously, a trade-off between the importance of fit and parameter reliability would be necessary to evaluate the cost/benefit of HML. No such trade-off is required here. The use of population-level information in establishing individual parameter estimates mitigates the estimation difficulties that ensnare MLE and bounded MLE, and consequently HML yields a better measure of individual differences.

Some caveats are in order, however. The benefits of hierarchical modeling may, for example, diminish when more choice data are available. In cases where many choices have been observed, the signal can more easily be teased apart from the noise and hierarchical methods may be less fruitful. Its benefits may also depend on the set of lotteries used. For lotteries specifically designed to test between models, the consequence being a narrow range of potential parameter values, the choices alone may be insufficient to reliably estimate individual risk preference parameters. Furthermore, cumulative prospect theory and the instance implemented here are but one particular choice model. The possibility exists that other models fit choice data better and that the resulting parameter estimates are even more reliable over time; in this case there may not be sufficient added value from hierarchical parameter estimation.

In this paper we have not discussed some alternative hierarchical estimation methods. Bayesian models have gained attention in the literature recently (Zyphur and Oswald 2013). A frequentist model is used here in part for its transparency and in part because it requires the addition of only one term in the likelihood function, given that the hierarchical distribution is known or can be reasonably approximated. Our data indicate that this distribution is fairly stable in a population and can be well represented with the log-normal function. Future results may indicate that this distribution can be applied “out of the box” to different contexts, which would make the HML method potentially attractive as a general tool for measuring risk preferences. The simple distributions used here stand in sharp contrast with the substantial flexibility offered by Bayesian methods. However, the “rigidity” of HML is (perhaps unexpectedly) valuable because it produces better out-of-sample performance and improved reliability. The simplicity of the hierarchical component of the HML method appears to prevent overfitting in-sample, which can undermine out-of-sample performance. Along these lines there is evidence of excessive shrinkage (Scheibehenne and Pachur 2013, 2015) with existing Bayesian hierarchical estimation methods when applied to cumulative prospect theory.

The hierarchical component that is used here is relatively simple. It is possible to modify both the hierarchical component as well as the extent to which the hierarchical method modulates the individual risk preference estimates, and there are certainly more complicated ways to implement the two-step procedure. One way to extend the HML method is to fit a new parameter that influences the weight of the hierarchical component by maximizing the fit in a retest, for example by truncating the probability of the hierarchical component to a lower bound and scaling it accordingly. Determining the advantages of such a method would require an experiment with three sessions: one to obtain parameter estimates, another to calibrate the hierarchical weighting parameters, and a third session to verify whether or not the calibrated model outperforms other methods. In this multitest setup, the correlation structure among parameters could also be estimated and used as an additional source of information in breaking “ties” in goodness of fit. As correlations among parameters exist, the use of partially dependent multivariate distributions may be useful. These extensions may be of both theoretical and practical interest, but are beyond the scope of the paper.

A major conclusion from the current paper is that the estimation method can greatly affect the inferences one makes about individual risk preferences, particularly in multiparameter risky choice models. A number of different parameter combinations may produce virtually the same fit result in-sample. When the interpretation of individual parameter values is used to make predictions about what an individual is like (e.g., psychological traits) or what she will do in the future (out-of-sample prediction), it is important to realize that other constellations of parameters are virtually equally plausible in terms of fit but may lead to vastly different conclusions about individual differences and predictions about behavior. This paper shows that using HML estimation methods can help us extract more “signal” from a set of noisy choices and thus yield a better measure of people’s innate risk preferences.

Acknowledgments

The authors thank Michael Schulte-Mecklenbeck, Thorsten Pachur, and Ralph Hertwig for the use of the experimental data analyzed in this paper. These data are a subset from a large decision-making study conducted at the Max Planck Institute for Human Development in Berlin. The authors also thank Helga Fehr-Duda, Thomas Epper, and Didier Sornette for helpful comments and suggestions on this project. This work was supported in part by the ETH Zürich Risk Center and is related to Future Resilient Systems (FRS) Module 3.1 from the Singapore ETH Centre.

Appendix A. Example of a Binary Risky Choice

A one-shot binary lottery (also referred to as a *gamble* or a *prospect*) is a common tool for studying risky decision making.[12] It is so common that it has been called the *fruit fly* of decision research (Lopes 1983). In these simple risky choices,

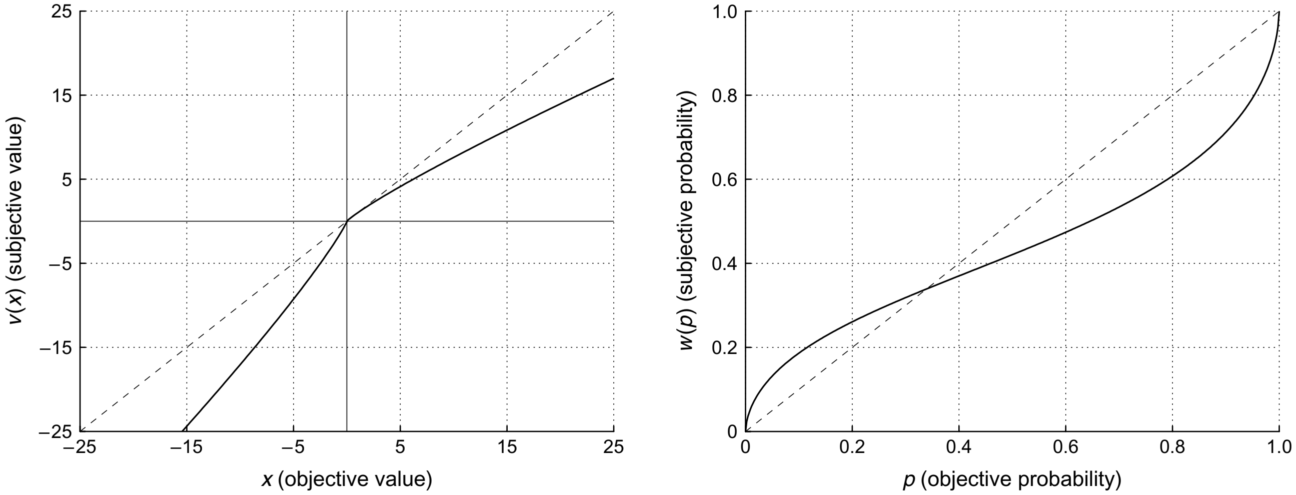
Table A.1. This Is an Example of a Simple, One-Shot Risky Decision Where the Choice Is Simply Whether to Select Option A or Option B

Option A			Option B		
$1	with probability	0.62	$37	with probability	0.41
$83		0.38	$24		0.59

Notes. This is one of the actual items used in the research, and the full list of items is included in Appendix E. The expected value of option A is higher ($32.16) than that of option B ($29.33), but option A contains the possibility of an outcome with a relatively small value ($1). As it turns out, the majority of DMs prefer option B, even though it has a smaller expected value.

Appendix B. Prospect Theory Curves

Figure B.1. A Typical Value Function (Left) and Probability Weighting Function (Right) from Prospect Theory



Notes. The parameters used to plot the solid lines are from Tversky and Kahneman (1992) and reflect the empirical findings at the aggregate level. The dashed lines represent risk-neutral preferences and veridical probability weighting. The solid line for the value function is concave over gains and convex over losses and further exhibits a "kink" at the reference point (in this case the origin) consistent with loss aversion. The probability weighting function overweights small probabilities and underweights large probabilities. It is worth noting that at the individual level, the plots can differ significantly from these aggregate-level plots.

DMs are presented with monetary outcomes x_i, each associated with an explicit probability p_i. Consider, for example, the lottery in Table A.1, offering a DM the choice between option A, a payoff of \$1 with probability 0.62 or \$83 with probability 0.38, or option B, a payoff of \$37 with probability 0.41 or \$24 with probability 0.59. The DM is called upon to choose either option A or option B and then the lottery can be played out via a random process for real consequences, thus ensuring incentive compatibility.

The expected value maximizing solution to the decision problem in Table A.1 is straightforward. Although this decision policy has desirable properties, it is often not an accurate description of what people prefer nor what they choose. Different people have different tastes of course (*de gustibus non est disputandum*), and this heterogeneity includes preferences for risk. For example, the majority of incentivized DMs select option B from the lottery shown in Table A.1, indicating, perhaps, that these DMs have some degree of risk aversion. However, the magnitude of this risk aversion is still unknown and it cannot be estimated from only one choice resulting from a binary lottery. This limitation has led researchers to use larger sets of binary lotteries where DMs make many independent choices; from the overall pattern of behavior, researchers can then draw inferences about the DM's underlying risk preferences. Preferences are revealed in this way, although the mapping from choices to model parameters is not so straightforward.

Appendix C. Interparameter Correlation

Table C.1 reports the interparameter correlations. The correlation between α (utility curvature) and φ (the noise parameter) is particularly low. This is not surprising because the value of α primarily affects utility (and scales its magnitude), whereas φ operates directly on the resulting utility.

Similar degrees of noise for lower values of α therefore lead to lower values of φ.[13] Other correlations are low or at most moderate. Note that it is not implausible that such correlations are real (i.e., not a measurement artifact), which may suggest some as yet undetermined (partial) common cause.

Appendix D. Aggregate Stochastic Predictions

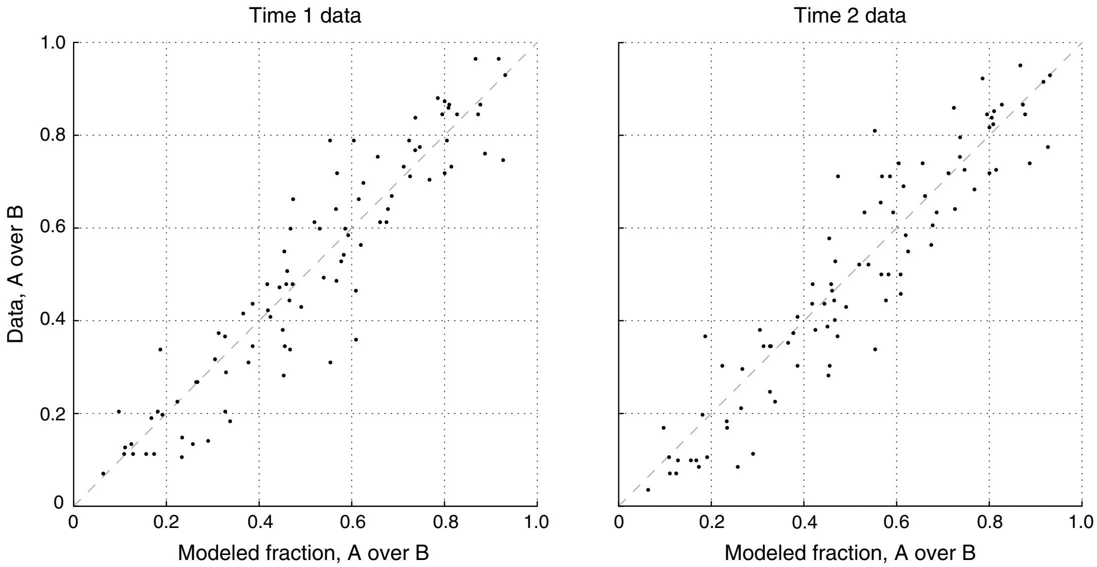
The workings of the stochastic choice function, shown in Equation (4), can be verified by comparing its predicted

Table C.1. Interparameter Correlation for MLE, Bounded MLE, and HML

	MLE					MLE bounded					HML				
	α	λ	δ	γ	φ	α	λ	δ	γ	φ	α	λ	δ	γ	φ
α	—	−0.10	0.07	−0.05	−0.62	—	−0.06	0.22	−0.04	−0.58	—	−0.21	−0.25	0.14	−0.44
λ		—	0.05	0.03	−0.10		—	−0.14	0.01	−0.15		—	−0.15	−0.06	−0.18
δ			—	0.27	−0.07			—	0.07	−0.23			—	0.11	0.12
γ				—	−0.11				—	−0.13				—	0.05
φ					—					—					—

Note. The correlation between α (utility curvature) and φ (noise parameters) for both MLE methods stands out.

Figure D.1. The Predicted Fraction of Subjects Choosing A over B vs. the Actual Fraction of Subjects Choosing A over B



Notes. Each dot represents one lottery. The correlation for both time 1 data (explanation) and time 2 data (prediction) is 0.93.

proportion of DMs selecting an option to the actually observed proportion of DMs choosing that option. This is shown in Figure D.1. For both time 1 as well as time 2 data, the correlation between the modeled fractions and the actual fractions is very high at 0.93.

Appendix E. Lotteries

Table E.1. Lotteries Used in the Experiment

									% that chose A	
Item	$p(A_1)$	A_1	$p(A_2)$	A_2	$p(B_1)$	B_1	$p(B_2)$	B_2	Session 1	Session 2
1	0.34	24	0.66	59	0.42	47	0.58	64	14	11
2	0.88	79	0.12	82	0.20	57	0.80	94	47	44
3	0.74	62	0.26	0	0.44	23	0.56	31	61	56
4	0.05	56	0.95	72	0.95	68	0.05	95	51	46
5	0.25	84	0.75	43	0.43	7	0.57	97	66	69
6	0.28	7	0.72	74	0.71	55	0.29	63	32	38
7	0.09	56	0.91	19	0.76	13	0.24	90	20	25
8	0.63	41	0.37	18	0.98	56	0.02	8	11	10
9	0.88	72	0.12	29	0.39	67	0.61	63	48	48
10	0.61	37	0.39	50	0.60	6	0.40	45	96	95
11	0.08	54	0.92	31	0.15	44	0.85	29	79	81
12	0.92	63	0.08	5	0.63	43	0.37	53	60	71
13	0.78	32	0.22	99	0.32	39	0.68	56	60	63
14	0.16	66	0.84	23	0.79	15	0.21	29	88	92
15	0.12	52	0.88	73	0.98	92	0.02	19	11	18
16	0.29	88	0.71	78	0.29	53	0.71	91	53	44
17	0.31	39	0.69	51	0.84	16	0.16	91	77	73
18	0.17	70	0.83	65	0.35	100	0.65	50	28	28
19	0.91	80	0.09	19	0.64	37	0.36	65	87	85
20	0.09	83	0.91	67	0.48	77	0.52	6	93	93
21	0.44	14	0.56	72	0.21	9	0.79	31	85	87
22	0.68	41	0.32	65	0.85	100	0.15	2	20	20
23	0.38	40	0.62	55	0.14	26	0.86	96	11	11
24	0.62	1	0.38	83	0.41	37	0.59	24	35	30
25	0.49	15	0.51	50	0.94	64	0.06	14	13	7
26	0.16	−15	0.84	−67	0.72	−56	0.28	−83	77	75
27	0.13	−19	0.87	−56	0.70	−32	0.30	−37	15	17
28	0.29	−67	0.71	−28	0.05	−46	0.95	−44	72	71
29	0.82	−40	0.18	−90	0.17	−46	0.83	−64	56	58
30	0.29	−25	0.71	−86	0.76	−38	0.24	−99	44	41

Table E.1. (Continued)

									% that chose A	
Item	$p(A_1)$	A_1	$p(A_2)$	A_2	$p(B_1)$	B_1	$p(B_2)$	B_2	Session 1	Session 2
31	0.60	−46	0.40	−21	0.42	−99	0.58	−37	96	92
32	0.48	−15	0.52	−91	0.28	−48	0.72	−74	70	68
33	0.53	−93	0.47	−26	0.80	−52	0.20	−93	46	50
34	0.49	−1	0.51	−54	0.77	−33	0.23	−30	73	72
35	0.99	−24	0.01	−13	0.44	−15	0.56	−62	79	84
36	0.79	−67	0.21	−37	0.46	0	0.54	−97	34	37
37	0.56	−58	0.44	−80	0.86	−58	0.14	−97	43	43
38	0.63	−96	0.37	−38	0.17	−12	0.83	−69	20	11
39	0.59	−55	0.41	−77	0.47	−30	0.53	−61	11	8
40	0.13	−29	0.87	−76	0.55	−100	0.45	−28	66	71
41	0.84	−57	0.16	−90	0.25	−63	0.75	−30	13	7
42	0.86	−29	0.14	−30	0.26	−17	0.74	−43	79	74
43	0.66	−8	0.34	−95	0.93	−42	0.07	−30	54	50
44	0.39	−35	0.61	−72	0.76	−57	0.24	−28	18	23
45	0.51	−26	0.49	−76	0.77	−48	0.23	−34	35	30
46	0.73	−73	0.27	−54	0.17	−42	0.83	−70	41	38
47	0.49	−66	0.51	−92	0.78	−97	0.22	−34	55	58
48	0.56	−9	0.44	−56	0.64	−15	0.36	−80	79	86
49	0.96	−61	0.04	−56	0.34	−7	0.66	−63	11	10
50	0.56	−4	0.44	−80	0.04	−46	0.96	−58	76	74
51	0.43	−91	0.57	63	0.27	−83	0.73	24	31	34
52	0.06	−82	0.94	54	0.91	38	0.09	−73	85	85
53	0.79	−70	0.21	98	0.65	−85	0.35	93	37	35
54	0.37	−8	0.63	52	0.87	23	0.13	−39	87	82
55	0.61	96	0.39	−67	0.50	71	0.50	−26	49	52
56	0.43	−47	0.57	63	0.02	−69	0.98	14	38	39
57	0.39	−70	0.61	19	0.30	8	0.70	−37	64	61
58	0.59	−100	0.41	81	0.47	−73	0.53	15	36	46
59	0.92	−73	0.08	96	0.11	16	0.89	−48	29	35
60	0.89	−31	0.11	27	0.36	26	0.64	−48	31	37
61	0.86	−39	0.14	83	0.80	8	0.20	−88	44	44
62	0.74	77	0.26	−23	0.67	75	0.33	−7	34	40
63	0.91	−33	0.09	28	0.27	9	0.73	−67	72	72
64	0.93	75	0.07	−90	0.87	96	0.13	−89	48	37
65	0.99	67	0.01	−3	0.68	74	0.32	−2	87	85
66	0.48	58	0.52	−5	0.40	−40	0.60	96	42	48
67	0.07	−55	0.93	95	0.48	−13	0.52	99	75	77
68	0.97	−51	0.03	30	0.68	−89	0.32	46	23	30
69	0.86	−26	0.14	82	0.60	−39	0.40	31	49	50
70	0.88	−90	0.12	88	0.80	−86	0.20	14	58	63
71	0.87	−78	0.13	45	0.88	−69	0.12	83	13	8
72	0.96	17	0.04	−48	0.49	−60	0.51	84	61	67
73	0.38	−49	0.62	2	0.22	19	0.78	−18	27	30
74	0.28	−59	0.72	96	0.04	−4	0.96	63	20	17
75	0.50	98	0.50	−24	0.14	−76	0.86	46	67	63
76	0.50	−20	0.50	60	0.50	0	0.50	0	73	73
77	0.50	−30	0.50	60	0.50	0	0.50	0	71	64
78	0.50	−40	0.50	60	0.50	0	0.50	0	70	55
79	0.50	−50	0.50	60	0.50	0	0.50	0	61	52
80	0.50	−60	0.50	60	0.50	0	0.50	0	48	44
81	0.50	−70	0.50	60	0.50	0	0.50	0	37	35
82	0.10	40	0.90	32	0.10	77	0.90	2	85	87
83	0.20	40	0.80	32	0.20	77	0.80	2	86	82
84	0.30	40	0.70	32	0.30	77	0.70	2	84	80
85	0.40	40	0.60	32	0.40	77	0.60	2	75	74
86	0.50	40	0.50	32	0.50	77	0.50	2	64	65
87	0.60	40	0.40	32	0.60	77	0.40	2	60	53
88	0.70	40	0.30	32	0.70	77	0.30	2	42	35
89	0.80	40	0.20	32	0.80	77	0.20	2	27	21
90	0.90	40	0.10	32	0.90	77	0.10	2	19	10
91	1	40	0	32	1	77	0	2	7	4

Notes. Each row is one lottery. The first column is the item number. The second through fifth columns describe option A, in the form of a $p(A_1)$ probability of A_1 and a $p(A_2)$ probability of A_2. The sixth through ninth columns describe option B in the same format. The last two columns list the fraction of subjects that chose option A over option B.

Endnotes

[1] Estimated parameter values for subjects who did not complete both sessions entirely did not differ significantly from subjects' parameters who did complete both sessions.

[2] Value x_1 (and its associated probability p_1) belongs to the option with the highest value for lotteries in the positive domain and to the option with the lowest associated value for lotteries in the negative domain. This ordering is to ensure that a cumulative distribution function is applied to the options systematically. See Tversky and Kahneman (1992) for details.

[3] It is possible to use $\alpha_{gain} \neq \alpha_{loss}$ by using an exponential utility function (Köbberling and Wakker 2005).

[4] Others (e.g., Abdellaoui 2000) have shown that the weighting function may differ between the domains of gains and losses. We do not dispute this. We use the same parameters for both domains for parsimony and as a first pass, given the relatively low number of binary observations compared to the number of model parameters.

[5] That point is $1/e \approx 0.368$ for $\delta = 1$, which is consistent with some empirical evidence (Gonzalez and Wu 1999, Camerer and Ho 1994). The point of intersection is sometimes also found to be closer to 0.5 (Fehr-Duda and Epper 2012) with aggregate data, which is consistent with Karmarkar's single-parameter weighting function (Karmarkar 1979). A two-parameter extension of that function is provided by Goldstein and Einhorn's function (Goldstein and Einhorn 1987).

[6] Furthermore, we consider only parameter values within $(0.01, 10)$, to prevent parameter estimates from running to more extreme values.

[7] It may be that a different approach with regard to the sensitivity parameter leads to better performance. We cannot judge this without data from a time 3 (a re-retest) because it requires estimation in a first session, calibration of the method using a second session, and true out-of-sample testing of this calibration in a third session. This may be interesting but is beyond the scope of the current paper.

[8] All resulting distribution parameters are roughly the same, but some differences do occur, of course, because of the use of random numbers, the use of the heuristic Nelder-Mead algorithm, and the use of different starting values. Note also that the resulting parameters are used in another parameter estimation step before individual-level parameters are obtained and that we have verified that small changes to the distribution parameters do not affect the pattern of results.

[9] Approximate values that can be used for the distributions, given by medians for estimates at time 1 and time 2 combined, are $D_\alpha \sim \mathrm{LN}(-0.31, 0.16)$, $D_\lambda \sim \mathrm{LN}(0.04, 0.64)$, $D_\delta \sim \mathrm{LN}(-0.15, 0.43)$, $D_\gamma \sim \mathrm{LN}(-0.33, 0.58)$.

[10] This is a consequence of fitting an asymmetric distribution to symmetric input data, potentially and easily circumvented by using a (bounded) normal distribution function for the hierarchical term instead.

[11] Deviance is a quality-of-fit statistic like the sum of squared residuals but is appropriate in cases where maximum likelihood is used rather than a least squares approach.

[12] See Harrison and Rutström (2008) for an extensive review of elicitation methods used in measuring risk preferences.

[13] Given the correlation between α and φ, a natural question is whether $\varphi = 1$ improves test-retest correlations. It does not. Test-retest correlations were significantly reduced as a result of fixing φ, except for α, which in the absence of a free φ takes on the role of both parameters.

References

Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. *Management Sci.* 46(11):1497–1512.

Barberis NC (2013) Thirty years of prospect theory in economics: A review and assessment. *J. Econom. Perspect.* 27(1):173–195.

Birnbaum MH (1999) The paradoxes of Allais, stochastic dominance, and decision weights. Shanteau J, Mellers BA, Schum DA, eds. *Decision Science and Technology* (Springer, New York), 27–52.

Birnbaum MH, Chavez A (1997) Tests of theories of decision making: Violations of branch independence and distribution independence. *Organ. Behav. Human Decision Processes* 71(2):161–194.

Bruhin A, Fehr-Duda H, Epper T (2010) Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica* 78(4):1375–1412.

Camerer CF (1995) Individual decision making. Kagel JH, Roth AE, eds. *The Handbook of Experimental Economics* (Princeton University Press, Princeton, NJ), 587–703.

Camerer CF (2004) Prospect theory in the wild: Evidence from the field. Camerer CF, Loewenstein G, Rabin M, eds. *Advances in Behavioral Economics* (Princeton University Press, Princeton, NJ), 148–161.

Camerer CF, Ho T-H (1994) Violations of the betweenness axiom and nonlinearity in probability. *J. Risk Uncertainty* 8(2):167–196.

Cavagnaro DR, Gonzalez R, Myung JI, Pitt MA (2013) Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management Sci.* 59(2):358–375.

Conte A, Hey JD, Moffatt PG (2011) Mixture models of choice under risk. *J. Econometrics* 162(1):79–88.

Edwards W (1954) The theory of decision making. *Psych. Bull.* 51(4):380–417.

Engelmann JB, Tamir D (2009) Individual differences in risk preference predict neural responses during financial decision-making. *Brain Res.* 1290:28–51.

Farrell S, Ludwig CJH (2008) Bayesian and maximum likelihood estimation of hierarchical response times. *Psychonomic Bull. Rev.* 15(6):1209–1217.

Fehr-Duda H, Epper T (2012) Probability and risk: Foundations and economic implications of probability-dependent risk preferences. *Ann. Rev. Econom.* 4(1):567–593.

Figner B, Murphy RO (2010) Using skin conductance in judgment and decision making research. Schulte-Mecklenbeck M, Kuehberger A, Ranyard R, eds. *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide* (Psychology Press, New York), 163–184.

Fox CR, Erner C, Walters DJ (2015) Decision under risk: From the field to the laboratory and back. *The Wiley Blackwell Handbook of Judgment and Decision Making* (John Wiley & Sons, Chichester, West Sussex, UK), 43–88.

Gächter S, Johnson EJ, Herrmann A (2007) Individual-level loss aversion in risky and riskless choice. IZA Discussion Paper 2961, Institute for the Study of Labor, Bonn, Germany.

Glöckner A, Pachur T (2012) Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition* 123(1):21–32.

Goldstein WM, Einhorn HJ (1987) Expression theory and the preference reversal phenomena. *Psych. Rev.* 94(2):236–254.

Gonzalez R, Wu G (1999) On the shape of the probability weighting function. *Cognitive Psych.* 38(1):129–166.

Harless DW, Camerer CF (1994) The predictive utility of generalized expected utility theories. *Econometrica* 62(6):1251–1289.

Harrison GW, Rutström EE (2008) Risk aversion in the laboratory. Cox J, Harrison G, eds. *Risk Aversion in Experiments*, Research in Experimental Economics, Vol. 12 (Emerald Group Publishing Limited, Bingley, UK), 41–196.

Harrison GW, Rutström EE (2009) Expected utility theory and prospect theory: One wedding and a decent funeral. *Experiment. Econom.* 12(2):133–158.

Hey JD, Orme C (1994) Investigating generalizations of the expected utility theory using experimental data. *Econometrica* 62(6): 1291–1326.

Holt CA, Laury SK (2002) Risk aversion and incentive effects. *Amer. Econom. Rev.* 92(5):1644–1655.

Huettel SA, Stowe CJ, Gordon EM, Warner BT, Platt ML (2006) Neural signatures of economic preferences for risk and ambiguity. *Neuron* 49(5):765–775.

Ingersoll J (2008) Non-monotonicity of the Tversky-Kahneman probability-weighting function: A cautionary note. *Eur. Financial Management* 14(3):385–390.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):263–291.

Kahneman D, Tversky A (2000) Choices, values and frames. Kahneman D, Tversky A, eds. *Choices, Values and Frames* (Cambridge University Press, Cambridge, UK), 1–16.

Karmarkar US (1979) Subjectively weighted utility and the Allais paradox. *Organ. Behav. Human Performance* 24(1):67–72.

Knight FH (1921) *Risk, Uncertainty, and Profit* (Houghton Mifflin Company, Chicago).

Köbberling V, Wakker PP (2005) An index of loss aversion. *J. Econom. Theory* 122(1):119–131.

Lewandowsky S, Farrell S (2010) *Computational Modeling in Cognition: Principles and Practice* (Sage, Los Angeles).

Lichtenstein S, Slovic P (1971) Reversals of preference between bids and choices in gambling decisions. *J. Experiment. Psych.* 89(1): 46–55.

Lichtenstein S, Slovic P (2006) *The Construction of Preference* (Cambridge University Press, Cambridge, UK).

Loomes G, Sugden R (1995) Incorporating a stochastic element into decision theories. *Eur. Econom. Rev.* 39(3):641–648.

Lopes LL (1983) Some thoughts on the psychological concept of risk. *J. Experiment. Psych. Human Perception Performance* 9(1):137–144.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (John Wiley & Sons, New York).

Luce RD, Raiffa H (1957) *Games and Decisions: Introduction and Critical Survey* (John Wiley & Sons, New York).

Luce RD, Suppes P (1965) Preference, utility, and subjective probability. Luce RD, Bush RR, Galante E, eds. *Handbook of Mathematical Psychology* (John Wiley & Sons, New York), 249–410.

McFadden D (1980) Econometric models for probabilistic choice among products. *J. Bus.* 53(3):S13–S29.

Mosteller F, Nogee P (1951) An experimental measurement of utility. *J. Political Econom.* 59(5):371–404.

Nelder JA, Mead R (1965) A simplex method for function minimization. *Comput. J.* 7(4):308–313.

Nilsson H, Rieskamp J, Wagenmakers EJ (2011) Hierarchical Bayesian parameter estimation for cumulative prospect theory. *J. Math. Psych.* 55(1):84–93.

Parducci A (1995) *Happiness, Pleasure, and Judgment: The Contextual Theory and Its Applications* (Psychology Press, New York).

Payne JW, Bettman JR, Johnson EJ (1992) Behavioral decision research: A constructive processing perspective. *Ann. Rev. Psych.* 43(1):87–131.

Pitt MA, Myung IJ (2002) When a good fit can be bad. *Trends in Cognitive Sci.* 6(10):421–425.

Prelec D (1998) The probability weighting function. *Econometrica* 66(3):497–527.

Rabin M (2000) Risk aversion and expected-utility theory: A calibration theorem. *Econometrica* 68(5):1281–1292.

Regenwetter M, Robinson M (2016) The construct-behavior gap in behavioral decision research: A challenge beyond replicability. Working paper, University of Illinois at Urbana–Champaign, Champaign.

Rieskamp J (2008) The probabilistic nature of preferential choice. *J. Experiment. Psych. Learn. Memory Cognition* 34(6):1446–1465.

Roberts S, Pashler H (2000) How persuasive is a good fit? A comment on theory testing. *Psych. Rev.* 107(2):358–367.

Samuelson PA (1938) A note on the pure theory of consumers' behaviour. *Economica* 5(17):61–71.

Scheibehenne B, Pachur T (2013) Hierarchical Bayesian modeling: Does it improve parameter stability? *Proc. 35th Annual Conf. Cognitive Sci. Soc., Berlin*, 1277–1282.

Scheibehenne B, Pachur T (2015) Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bull. Rev.* 22(2):391–407.

Schulte-Mecklenbeck M, Pachur T, Murphy RO, Hertwig R (2016) Prospect theory tracks selective allocation of attention. Working paper, Max Planck Institute, Berlin.

Starmer C (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *J. Econom. Literature* 38(2):332–382.

Stevens SS (1957) On the psychophysical law. *Psych. Rev.* 64(3): 153–181.

Stewart N, Chater N, Brown GDA (2006) Decision by sampling. *Cognitive Psych.* 53(1):1–26.

Stewart N, Chater N, Stott HP, Reimers S (2003) Prospect relativity: How choice options influence decision under risk. *J. Experiment. Psych.: General* 132(1):23–46.

Stott HP (2006) Cumulative prospect theory's functional menagerie. *J. Risk Uncertainty* 32(2):101–130.

Tanaka T, Camerer CF, Nguyen Q (2010) Risk and time preferences: Linking experimental and household survey data from Vietnam. *Amer. Econom. Rev.* 100(1):557–571.

Toubia O, Johnson EJ, Evgeniou T, Delquié P (2013) Dynamic experiments for estimating preferences: An adaptive method of eliciting time and risk parameters. *Management Sci.* 59(3):613–640.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representations of uncertainty. *J. Risk Uncertainty* 5(4): 297–323.

Varian HR (2006) Revealed preference. Szenberg M, Ramrattan L, Gottesman AA, eds. *Samuelsonian Economics and the Twenty-First Century* (Oxford University Press, Oxford, UK), 99–115.

Wakker PP (2005) Formalizing reference dependence and initial wealth in Rabin's calibration theorem. Working paper, Erasmus University, Rotterdam, Netherlands.

Wakker PP (2010) *Prospect Theory: For Risk and Ambiguity* (Cambridge University Press, Cambridge, UK).

Wetzels R, Vandekerckhove J, Tuerlinckx F, Wagenmakers EJ (2010) Bayesian parameter estimation in the expectancy valence model of the Iowa gambling task. *J. Math. Psych.* 54(1):14–27.

Willemsen MC, Johnson EJ (2011) Visiting the decision factory: Observing cognition with MouselabWEB and other information acquisition methods. Schulte-Mecklenbeck M, Kühberger A, Ranyard R, eds. *A Handbook of Process Tracing Methods for Decision Research: A Critical Review and User's Guide* (Taylor and Francis, New York), 21–42.

Willemsen MC, Johnson EJ (2014) Mouselab software. Last accessed November 17, 2016, http://www.mouselabweb.org.

Zeisberger S, Vrecko D, Langer T (2012) Measuring the time stability of prospect theory preferences. *Theory Decision* 72(3):359–386.

Zyphur MJ, Oswald FL (2013) Bayesian probability and statistics in management research: A new horizon. *J. Management* 39(1):5–13.