Decision
2014, Vol. 1, No. 4, 252–274

© 2014 American Psychological Association
2325-9965/14/$12.00 http://dx.doi.org/10.1037/dec0000020

Parameter Recovery for Decision Modeling Using Choice Data

Stephen B. Broomell
Carnegie Mellon University

Sudeep Bhatia
University of Warwick

We introduce a general framework to predict how decision sets used in decision-making experiments impact the quality of parameter estimates. We applied our framework to cumulative prospect theory (CPT) to investigate the expected parameter discrimination achieved by current research practices. Our approach revealed several regularities in the ability of recent decision sets to recover CPT parameters. We analyzed 3 decision sets and we found that (a) the randomly generated stimuli performed just as well as the researcher designed stimuli, (b) outcome magnitude impacted the recoverability of parameters, and (c) even under ideal circumstances, the parameters representing loss aversion and choice sensitivity were associated with large amounts of error in their estimates. Our analysis is the first to accurately predict the relative estimation precision of each parameter of CPT. We additionally applied our framework to analyze the decision sets that were used to produce the empirical evidence for the description–experience gap. Specifically, we found that choices based on few experienced draws from a gamble provided little information for estimating decision weights when compared to equivalent description based choices. Therefore, choices between experienced gambles can be explained by a wider range of decision weights than choices between equivalent described gambles, providing an alternative explanation for the empirical evidence surrounding the description–experience gap. We conclude with implications for future experiments designed to estimate parameters from choice data.

Keywords: choice data, parameter estimation, experimental design, decision modeling

Supplemental materials: http://dx.doi.org/10.1037/dec0000020.supp

We introduce a general mathematical framework to investigate the effectiveness of decision sets to estimate free parameters from any cognitive model that produces a probabilistic prediction of choice. For example, extensive research has been dedicated to estimating the free parameters for decision models that include cumulative prospect theory (CPT; Tversky & Kahneman, 1992), decision field theory (Busemeyer & Townsend, 1993), and models of time discounting (Frederick, Loewenstein, & O'Donoghue, 2002). While our approach could be applied to any psychological domain that predicts dichotomous choices, our current focus is on monetary decision-making research.

The goal of many decision-making experiments is to infer decision-makers' (DMs') preferences by observing their choices over a set of decision problems. Those underlying preferences are typically represented by the free parameters of cognitive decision models estimated from these choices. To achieve accurate estimates, the collection of decision problems (henceforth *decision set*) must be designed so the observed choice data can be used to distinguish between individuals with different preferences (i.e., different parameter values). Although parameter estimation from choice data is common practice in decision-making research, the expected power of a decision set to discriminate between DMs with different parameters is not usually considered when evaluating experimental designs. We therefore propose a framework to measure the ability of choice data to reveal an individual's true preferences.

Stephen B. Broomell, Department of Social and Decision Sciences, Carnegie Mellon University; Sudeep Bhatia, Warwick Business School, University of Warwick.

Special thanks to Baruch Fischhoff, Russell Golman, Chad M. Schafer, and the blind reviewers for their helpful comments and suggestions on earlier drafts.

Correspondence concerning this article should be addressed to Stephen B. Broomell, Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: broomell@gmail.com

Measuring the ability of a decision set to distinguish between potential parameter values has many benefits for guiding experimental design. First, it can be used to facilitate the creation of more efficient experiments. For example, researchers can remove decision problems that are uninformative by identifying individual decision problems (or sets of decision problems) where predicted choices do not depend on the parameters. Additionally, decision models often have multiple parameters of interest, but the relative estimation precision of these parameters is unknown. This is especially important because differences in the precision of estimated parameters (constrained to a bounded range) can systematically bias conclusions drawn from different decision sets (e.g., see our analysis of the description–experience gap).

We demonstrate the desirability of our approach by using it to predict parameter recovery for CPT (Tversky & Kahneman, 1992). Of the three decision sets we analyzed, we found that (a) randomly generated stimuli performed just as well as researcher designed stimuli, (b) outcome magnitude impacted the recoverability of parameters, and (c) the parameters such as loss aversion and choice sensitivity were associated with large amounts of error in their estimates. Finally, our approach also revealed a potential problem with the empirical analysis of the description–experience gap (D–E gap; Hau, Pleskac, Kiefer, & Hertwig, 2008; Hertwig, Barron, Weber, & Erev, 2004). Decision sets related to the D–E gap may suffer from a design problem that reduces their ability to accurately reveal the impact of probability on choice. We were able to contribute to these topics by investigating the statistical properties of experimental designs currently used for studying prospect theory.

Methods for Parameter Estimation

There are two prominent methods for estimating decision model parameters in psychological research that differ in respect to estimation accuracy: *direct scaling* and *choice data estimation*. One example of direct scaling relies on eliciting dollar amounts that are equal in subjective value to a gamble (i.e., a certainty equivalent; Gonzalez & Wu, 1999; Luce, 1992; Prelec, 1998; Tversky & Fox, 1995). Choice data estimation relies on the elicitation of choices between two or more gambles in order to back-calculate underlying preferences (Glöckner & Pachur, 2012; Nilsson, Rieskamp, & Wagenmakers, 2011).

While direct scaling has been heavily relied on in psychological research, there are several limitations to this method. First, scaling can be time consuming. Some elicitation procedures require personal interaction between the decision analyst and the DM in order to ensure mutual understanding of the procedures. Second, scaling results can differ depending on the method of elicitation that is used by the researcher (Lichtenstein & Slovic, 1971; Por & Budescu, 2013; Tversky, Slovic, & Kahneman, 1990; Tversky & Thaler, 1990) and the DM's ability to express their values in the required response mode (Fischhoff, 1991; Fischhoff, 2005). One potential cause for these differences is that bias is introduced by scaling methods. Indeed, recent studies have demonstrated bias in the elicitation of certainty equivalents due to context and mere-presentation effects (Erev, Glozman, & Hertwig, 2008).

Using choice data estimation as an alternative approach to estimating decision model parameters has become more prominent in recent years. Choice data is commonly used to find best fit decision model parameters, typically with numerical maximum likelihood methods (Glöckner & Pachur, 2012; Pachur, Hanoch & Gummerum, 2010; Rieskamp, 2008; Stott, 2006).[1] While the elicitation of choices can reduce participant biases, it can also have the unintended consequence of increasing the variance in estimated parameters.

Parameter Recovery From Choice Data

Optimizing parameter recovery for decision-making experiments involves evaluating the decision problems in terms of the information each response provides regarding the parameters of interest. Information in most choice based experiments is defined using Fisher information, and experimental designs that maximize this metric of information are considered D-optimal (Kiefer, 1959). D-optimal designs for nonlinear decision models depend on the

[1] See Nilsson et al. (2011) for alternative estimation methods such as hierarchical Bayesian estimation.

true values of the parameters being estimated. When these values are unknown, determining a single optimal design becomes more difficult.

Chang and Ying (1996) propose an approach based on Kullback-Leibler (KL) divergence that is particularly useful in these settings. They apply KL divergence to the likelihood functions given by two competing parameter values for a set of decision problems, in order to measure the ability of the decision problems to discriminate between the two parameter values. Chang and Ying show that designing experiments using KL divergence is superior to designing stimuli using Fisher information when there is large uncertainty about the true parameter value.

An alternative framework for maximizing parameter estimation uses the Bayes D-optimum criterion. This approach has also been extensively applied to adaptive design optimization for cognitive and decision models (Cavagnaro, Gonzalez, Myung, & Pitt, 2012; Cavagnaro, Pitt, & Myung, 2011; Myung, Cavagnaro, & Pitt, 2013). This criterion optimizes the reduction in uncertainty about the true parameter values achieved by an experiment. However, Bayes D-optimality is computationally complex, requiring the evaluation of the joint information gain from an entire experiment. For a choice experiment with N decision problems, this requires 2^N computations inside an integral across the prior parameter distribution.

While both of these frameworks would work for our purposes, we use the criteria introduced by Chang and Ying (1996) and extended by Veldkamp and van der Linden (2002). This criteria was developed for adaptive designs that optimize the next stimuli based on previous responses. We have adapted this criterion to develop a framework for evaluating static designs where a single set of stimuli is given to each DM. Static experiments typically use maximum likelihood estimation for parameter recovery, which requires all the assumptions needed for Chang and Ying's criterion, directly linking the criterion's interpretation with the estimation routine. We use this information criterion as a measure to rank decision sets for what we will call *parameter discrimination*.[2] This measure can be interpreted as the expected power of a static decision set to distinguish a DM from a different DM, when both DMs are independently and randomly drawn from a distribution of DMs. This interpretation perfectly matches the setup of a majority of laboratory experiments designed to use a single decision set to estimate parameters from a collection of randomly sampled DMs.

Our proposed methodology can be used to (a) evaluate the effectiveness of existing decision sets, (b) evaluate the impact of experimental manipulations on parameter estimation, and (c) compare the relative precision of multiple parameters. In addition to demonstrating the usefulness of this approach for choice experiments, we outline the relationship between expected parameter discrimination and the Bayes D-optimal criteria for static experimental designs (also see Wang & Chang, 2011, for comparisons within computerized adaptive testing). Specifically, we show that our criterion is an upper bound for the Bayes D-optimality criteria. Unlike previous research on (adaptive) design optimization, our approach is less computationally intensive and does not require previous (or any) responses from a DM. We are not proposing a substitute for adaptive design or design optimization. However, we are proposing a framework for evaluating parameter recovery that can be easily computed and interpreted by behavioral decision researchers using static experimental designs.

A Measure of Expected Parameter Discrimination

Measuring Divergence of the Likelihood Function

Consider a set of stimuli consisting of N binary decision problems that are used to elicit N choices from a DM.[3] The vector of choices is denoted $\boldsymbol{x} = [x_1 \ldots x_i \ldots x_N]$ with $x_i = 1$ if the first option is chosen, and $x_i = 0$ if the second option is chosen. Let $\boldsymbol{\theta}$ represent the vector of parameters for a particular cognitive decision model that predicts a probability for choosing option 1 denoted $\pi_i(\boldsymbol{\theta})$. The likelihood of a particular parameter vector given the observed choices is defined as

[2] We introduce this terminology because the term "information" is interpreted differently by the two standard approaches to optimal designs.

[3] In this paper, we are considering only binary decision problems. Our framework, however, trivially generalizes to the settings with three or more decision problems.

$$L(\boldsymbol{\theta} \mid x_1, \ldots, x_N) = p(\mathbf{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \pi_i(\boldsymbol{\theta})^{x_i}[1 - \pi_i(\boldsymbol{\theta})]^{1-x_i}. \quad (1)$$

Next we compute the KL divergence between two likelihood functions assuming different parameter values. Measures of KL divergence are expressed as $D_{KL}(p(A) \| p(B))$, where the double bars indicate divergence, and defined as the expected log-likelihood of the two distributions with respect to the first distribution, given by

$$D_{KL}(p(A) \parallel p(B)) = E_{p(A)} \log\left[\frac{p(A)}{p(B)}\right]. \quad (2)$$

The KL divergence between any two distributions is non-negative and can be asymmetric, meaning that $D_{KL}(p(A) \| p(B)) \geq 0$ need not equal $D_{KL}(p(B) \| p(A)) \geq 0$. KL divergence is also a measure of relative entropy (Cover & Thomas, 1991; Kullback & Leibler, 1951). Entropy in *bits* is computed with logarithms of base 2 and can be interpreted as the average number of yes/no questions required to describe a distribution. Relative entropy is defined as the additional bits of information required to describe the true distribution when assuming an alternative distribution.

Let a DM's true parameters be denoted as $\boldsymbol{\theta}_0$ and alternative parameters be denoted as $\boldsymbol{\theta}_1$. The parameter discrimination for an entire experiment can be expressed as

$$D_{KL}(p(\mathbf{x} \mid \boldsymbol{\theta}_0) \| p(\mathbf{x} \mid \boldsymbol{\theta}_1)) = \sum_{i=1}^{N} D_{KL}(p(x_i \mid \boldsymbol{\theta}_0) \| p(x_i \mid \boldsymbol{\theta}_1)). \quad (3)$$

This value is a measure of the ability of the entire decision set to distinguish the true parameter vector ($\boldsymbol{\theta}_0$) from the alternate parameter vector ($\boldsymbol{\theta}_1$). The parameter discrimination of the decision set is equal to the sum of the parameter discrimination for each decision problem. Because KL divergence is non-negative, each additional decision problem will monotonically increase the discriminal ability of the decision set. Next we outline a series of aggregate level measures, each based on more restrictive assumptions: (a) expected multivariate parameter discrimination (MPD), (b) expected univariate parameter discrimination (UPD), and (c) parameter discrimination as an effect size.

Multivariate parameter discrimination. The expected divergence between likelihoods given competing parameters is calculated by modeling the uncertainty in the parameters using a distribution $p(\boldsymbol{\theta})$ over the parameter space, adopting a Bayesian philosophy with $p(\boldsymbol{\theta})$ as the prior (Veldkamp & van der Linden, 2002). We define MPD as the expected value of the divergence between the likelihood functions given $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ expressed as

$$\begin{aligned} MPD_{\boldsymbol{\theta}} &= E_{p(\boldsymbol{\theta}_0)} E_{p(\boldsymbol{\theta}_1)}\left[D_{KL}(p(\mathbf{x} \mid \boldsymbol{\theta}_0) \| p(\mathbf{x} \mid \boldsymbol{\theta}_1))\right] \\ &= \int_{\boldsymbol{\theta}_0}\int_{\boldsymbol{\theta}_1} D_{KL}(p(\mathbf{x}|\boldsymbol{\theta}_0) \| p(\mathbf{x}|\boldsymbol{\theta}_1)) p(\boldsymbol{\theta}_0) \\ &\quad \times p(\boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_0. \end{aligned} \quad (4)$$

Assuming that the prior distributions for the true and alternative parameters are interchangeable, this expression produces a symmetrized measure of divergence. The prior distributions can be thought of as weighting functions for computing the aggregate measure. Any theorized probability distribution can be used for $p(\boldsymbol{\theta})$ that has support over the range of viable parameter values, for example, the uniform distribution will compute a simple mean (see the Applications section for more details).

Additionally, MPD is an upper bound for the Bayes D-optimum criterion (see Appendix A for proof),

$$\begin{aligned} &\text{Bayes D-optimum} \\ &\quad = E_{\mathbf{x}}[D_{KL}(p(\boldsymbol{\theta}_0 \mid \mathbf{x}) \| p(\boldsymbol{\theta}_0))] \\ &\quad = E_{p(\boldsymbol{\theta}_0)}[D_{KL}(p(\mathbf{x} \mid \boldsymbol{\theta}_0) \| p(\mathbf{x}))] \\ &\quad \leq E_{p(\boldsymbol{\theta}_0)} E_{p(\boldsymbol{\theta}_1)}[D_{KL}(p(\mathbf{x} \mid \boldsymbol{\theta}_0) \| p(\mathbf{x} \mid \boldsymbol{\theta}_1))]. \end{aligned} \quad (5)$$

MPD can be interpreted as the expected power to discriminate between the likelihood of two randomly selected parameters whereas the Bayes D-optimum criterion can be interpreted as the expected uncertainty reduction between of the likelihood of the true parameter and the marginal likelihood across all potential parameter values.[4] Based on this relationship, we can draw the following conclusions: (a) low measures of MPD, such as an MPD = 0, indicate that parameters cannot be estimated with the current experiment; and (b) decision sets with high MPD can be interpreted as having good power, on average, to discriminate a randomly drawn DM from another randomly drawn DM, assuming each choice is made independently. The limitations of MPD are the following: (a) maximizing MPD does not imply Bayesian optimality, and (b) MPD cannot jointly account for the information obtained from the experiment; each decision problem is treated as an independent experiment. However, the assumption of independence is also required for maximum likelihood estimation and the simplicity of MPD allows easier computation when studying nonlinear cognitive models with many decision problems. While the Bayes D-optimum criterion may be better at optimizing decision set efficiency for each individual, the proposed MPD measure presents a desirable approach to analyzing static experiments designed for maximum likelihood estimation.

Univariate parameter discrimination. Next, we outline a method for decomposing the MPD measure in Equation 4 to each individual model parameter. The first integral in Equation 4 is computed over the true parameter values, $\boldsymbol{\theta}_0$, and the second integral is computed over the alternative parameter values, $\boldsymbol{\theta}_1$. We can compute Equation 4 for a subset of parameter values by constraining any of the individual alternative parameters to be equal to the true parameter value. This operation will reduce the dimensions of the alternative parameter integral by each constrained parameter. For example, consider two parameters such that $\boldsymbol{\theta} = [\alpha, \gamma]$. Assuming the true value of γ is known, we can compute Equation 4 by constraining $\gamma_1 = \gamma_0$ as follows:

$$MPD_{\alpha} = \int_{\boldsymbol{\theta}_0} \left[\int_{\alpha_1} D_{KL}(p(\boldsymbol{x} \mid \alpha_0, \gamma_0) \parallel p(\boldsymbol{x} \mid \alpha_1, \gamma_0)) \times p(\alpha_1) d\alpha_1 \right] p(\boldsymbol{\theta}_0) d\boldsymbol{\theta}_0. \tag{6}$$

The subscript denotes which parameters are unconstrained. In the two-parameter case, this measure computes the expected discrimination for the parameter α assuming that the true value of γ is known. However, this measure for discriminating α does not account for any potential reduction in the discrimination of α due to the misspecification of γ.

To account for potential parameter interactions, we develop a measure that accounts for the expected discrimination produced by a target parameter in the presence of the remaining parameters. For any alternative parameter vector with J model parameters, there are 2^J total ways to constrain subsets of parameters. We can compute the *marginal divergence* of a target parameter by subtracting the MPD for a subset of parameters where a target parameter is constrained from the MPD for the same parameter subset where the target parameter is unconstrained. We propose a measure of UPD as the overall average of the marginal divergence of a target parameter across all parameter subsets. This measure provides a useful decomposition of the MPD to each of the individual parameters and accounts for the interactions between parameters by considering all possible parameter combinations. For example, the UPD of α in the two-parameter example is given by

$$UPD_{\alpha} = .5(MPD_{[\alpha,\gamma]} - MPD_{\gamma}) + .5(MPD_{\alpha} - 0). \tag{7}$$

This measure is the average difference in MPD when α is constrained to be known compared to when α is unconstrained across all parameter combinations with constrained and unconstrained γ. The interpretation of UPD is the same as MPD. When UPD = 0, the target parameter cannot be estimated with the current experiment. Higher UPD predicts better power, on average, to discriminate between two randomly drawn values of the target parameter.

Additionally we can also derive a measure of parameter interaction. We define the percent reduced discrimination (PRD) for the target parameter α as

$$PRD_{\alpha} = 1 - (UPD_{\alpha}/MPD_{\alpha}). \tag{8}$$

[4] We thank a blind reviewer for pointing out the connection between these two approaches.

This measure can be computed for each parameter in the decision model and is interpreted as a measure of parameter interaction. More specifically, this measure captures the percent reduction in expected parameter discrimination caused by the misspecification of the remaining parameters in the model. This measure is bounded by the interval (0, 1) and a PRD of 0% indicates that the misspecification of the remaining parameters has no impact on recovering the target parameter. Larger values of PRD indicate that the remaining parameters have more influence on the recovery of the target parameter.

Interpreting Parameter Discrimination as an Effect Size

The measures presented above are useful for comparing two or more decision sets (we demonstrate the effectiveness of these measures to rank existing decision sets in the next section). While we believe these measures are valuable complements to existing statistical tools in model fitting, experimenters who use these measures still face an important problem: how to evaluate whether a particular decision set is good enough.

Such problems are ubiquitous in the experimental sciences. While it is fairly easy to compare the relative strength of two phenomena, or in our case, the relative desirability of two decision sets, it is difficult to establish objective guidelines to evaluate an individual effect size for a decision set independently of a comparison standard. One solution to this problem is to use existing consensus. For example, signal detection theory (Green & Swets, 1966) provides a measure of discrimination ability, d', which is a standardized measure of the distance between two normal distributions (similar to the effect size measure Cohen's d; Cohen, 1988). It is possible to translate parameter discrimination into the domain of signal detection theory to facilitate interpretation of the overall adequacy of a decision set.

Univariate parameter discrimination effect size. The measure d' is a mean shift between two standard normal distributions, and we can analytically solve for the KL divergence for such a shift. Let $p(x) = N(\mu_0, 1)$ and $q(x) = N(\mu_0 + d', 1)$, then we can solve for $D_{bits} = D_{KL}(p(x) \| q(x))$ as the following equality:

$$D_{bits} = \frac{(d')^2}{2 * \ln(2)}. \tag{9}$$

The KL divergence is a function of the mean shift d' and does not depend on the value of μ_0 (see proof in Appendix B). Finally, we solve for d' so the mean shift between two standard normal distributions can be computed as a function of the KL divergence, given by

$$d' = \sqrt{2 * \ln(2) * D_{bits}}. \tag{10}$$

This equality shows that an estimate of KL divergence between any two distributions (denoted D_{bits}) can be easily related to an equivalent d' for a Gaussian signal detection experiment. We propose evaluating the discriminability of parameters estimated from choice data by computing an equivalent d' based on the expected discrimination generated by a constant shift in a parameter value. This measure can then be evaluated based on the same (existing) standards as those applied to signal detection theory for d' (or effect sizes using Cohen's d).

In order to use the signal detection framework we need to alter our measures to better mimic the divergence of normal distributions. The divergence between two normal distributions has two unique properties associated with the location scale family: (a) It is symmetric, that is $D_{KL}(N(0, 1) \| N(d', 1)) = D_{KL}(N(d', 1) \| N(0, 1))$. (b) The computation is a function only of d' and not of μ_0. These properties do not hold in more general settings. To account for these differences, our effect size measure is based on a symmetrized measure of KL divergence. The parameter discrimination effect size for a target parameter is computed using the same method as UPD, except the alternative target parameter is constrained to be the true parameter plus a shift δ. For example, using the two-parameter case where $\boldsymbol{\theta} = [\alpha, \gamma]$, discriminating α parameters that are a distance of δ apart from each other is computed using the following symmetrized measure of KL divergence:

$$\frac{1}{2}\Big[D_{KL}(p(\mathbf{x} \mid \alpha_0, \gamma_0) \| p(\mathbf{x} \mid \alpha_0 + \delta, \gamma_1)) + D_{KL}(p(\mathbf{x} \mid \alpha_0 + \delta, \gamma_1) \| p(\mathbf{x} \mid \alpha_0, \gamma_0))\Big] \tag{11}$$

The expected effect size in bits, denoted $\hat{D}(\delta)$, is computed by replacing the KL divergence function in Equation 6 with the symmetrized KL divergence function in Equation 11. The effect size for the discrimination of each parameter is computed using the exact same method outlined for computing UPD (as the average marginal divergence of the target parameter shifted by δ). The value of $\hat{D}(\delta)$ can be used to estimate $\hat{d}' = \sqrt{2 * \ln(2) * \hat{D}(\delta)}$ as the distance between two normal distributions, which is equivalent to the expected divergence of the target parameter by a shift of δ.

The interpretation of $\hat{d}'$ is the number of standard deviations between the means of two standard normal distributions. We can judge the usefulness of a given $\hat{d}'$ by the expected false alarms and correct hits generated by a decision threshold that symmetrically divides the two distributions. The false alarm rate and hit rate for $\hat{d}' = 1$ are 0.31 and 0.69, respectively. An estimate of $\hat{d}' \geq 4$ indicates very good discriminal ability (false alarm rate < 0.05; hit rate $>$ 0.95) and $\hat{d}' \geq 6$ indicates nearly error free discriminal ability (false alarm rate < 0.001; hit rate > 0.999).

Applications

We apply our methodology to CPT to demonstrate the flexibility of our approach. CPT is the most widely used model of risky choice and contains many complicated mathematical operations that limit the ability to solve for optimal designs. Our methodology and computational approach is performed numerically and can therefore be easily extended to any cognitive decision model (with free parameters) that produces a probabilistic prediction, such as decision field theory (Busemeyer & Townsend, 1993) or configural-weight theory (Birnbaum, 2008) for risky choice; exponential, hyperbolic, or quasihyperbolic discounting (Frederick et al., 2002) for intertemporal choice; and inequality aversion (Fehr & Schmidt, 1999) for altruistic choice.

Cumulative Prospect Theory's Functional Forms

Risky decision-making involves the evaluation of gambles. We can write a two-outcome gamble G, with each outcome y_j having a p_j chance of being realized, as $(y_1, p_1; y_2, p_2)$. Decision models for risky choice are often compared using choices between these gambles. For example, Kahneman and Tversky's (1979) seminal research used 12 decision problems with each problem consisting of two such gambles. Kahneman and Tversky used differences in observed choice proportions to demonstrated the descriptive superiority of prospect theory over expected utility. More recently, it has become standard practice to use choice data over many pairs of gambles to estimate individual-level or pooled group-level parameters for different risky decision models (see, Donkers, Melenberg, & van Soest, 2001; Glöckner & Pachur, 2012; Harrison, Humphrey, & Verschoor, 2009; Holt & Laury, 2002; Nilsson et al., 2011; Pachur et al., 2010; Rieskamp, 2008; Stott, 2006; Wu & Markle, 2008).

CPT (Tversky & Kahneman, 1992) assumes that DMs display diminishing sensitivity in both gains and losses, overweigh small probabilities, and underweigh large probabilities. Many parameterizations and functional forms for CPT have been developed and tested over the years. Following the recommendation of Stott (2006), we model CPT using the power value function (Tversky & Kahneman, 1992) and the Prelec (1998) one-parameter probability weighting function. As we only consider choices between pairs of gambles, we use the logit choice function (also recommended by Stott) to transform gamble values into choice probabilities. In the following paragraphs, we introduce a vector of 4 parameters $\boldsymbol{\theta} = [\alpha, \gamma, \lambda, \varepsilon]$ that captures CPT based on these functional forms.

The probabilistic choice prediction from CPT is given by a logit transformation:

$$\pi(G_1) = \frac{1}{1 + e^{-\varepsilon\,(V(G_1) - V(G_2))}}. \qquad (12)$$

The choice probability is a function of the *choice-sensitivity parameter*, ε, and the subjective values of each gamble, $V(G_1)$ and $V(G_2)$. The choice-sensitivity parameter determines the amount of randomness in an individuals' choice, where $\varepsilon = 0$ implies completely random choices regardless of subjective value. Increasing values of ε imply more correspondence to differences in subjective value.

The subjective evaluations of gambles follow the CPT (Tversky & Kahneman, 1992) framework of separating gains and losses, given by

$$V(G) = \sum_{+} v(y_j) * w^{+}(p_j) + \sum_{-} v(y_j) * w^{-}(p_j). \tag{13}$$

The value of a gamble, $V(G)$, is the combined valuation of the positive (+) and negative (−) outcomes in a gamble, defined as the sum of the subjective value of each outcome, $v(y_j)$, multiplied by the decision weight, $w(p_j)$. The (+/−) symbols identify the gain/loss separation for computing cumulative probability weights. The subjective value function is given by

$$v(y_j) = \begin{cases} y_j^{\alpha} & \text{if } y_j \geq 0 \\ -\lambda(-y_j)^{\alpha} & \text{if } y_j < 0 \end{cases}. \tag{14}$$

The parameter α determines the sensitivity to changes in value such that $\alpha < 1$ produces diminishing sensitivity. The parameter λ determines loss aversion such that $\lambda > 1$ produces choices where losses loom larger than gains.

The probability weighting function is given by

$$W^{(+/-)}(p_j) = e^{-(-\ln(p_j))^{\gamma}}. \tag{15}$$

The parameter γ determines the decision weights such that $\gamma < 1$ produces choices that overweight small probabilities.

The decision weights are computed by separating positive and negative outcomes and computing the difference between cumulative weights separately for gains and losses. Let outcomes be ordered from smallest to largest such that $y_{j-1} < y_j < y_{j+1}$. The weight of outcome y_j having probability of p_j is given by

$$w^{+}(p_j) = W^{+}(p_j \cup p_{j+1}) - W^{+}(p_{j+1}) \text{ if } y_j > 0, \tag{16}$$

$$w^{-}(p_j) = W^{-}(p_{j-1} \cup p_j) - W^{-}(p_{j-1}) \text{ if } y_j < 0. \tag{17}$$

The subjective evaluation of each gamble is a function of the diminishing-sensitivity parameter, α, the probability weighting parameter, γ, and the loss-aversion parameter, λ. The negative and positive outcomes are constrained to have the same diminishing-sensitivity parameter, α, and probability weighting parameter, γ. This ensures that the only difference between losses and gains are produced by the loss-aversion parameter λ. Nilsson et al. (2011) recommend constraining risk tolerance parameters to be equal for gains and losses to achieve better estimates of loss aversion, λ, and reduce parameter interaction. We also constrained the decision-weighting parameter in the same way to simplify our analysis.

Parameter Discrimination for Cumulative Prospect Theory

Our proposed measures of parameter discrimination can be used to quantitatively evaluate experimental decision sets. We selected three decision sets recently used for CPT parameter estimation referred to as Stott (2006); Erev, Roth, Slonim, and Barron (2002); and Glöckner and Pachur (2012). We included these decision sets because they were all used to estimate parameters from CPT using maximum likelihood, but differed with respect to many dimensions: (a) the total number of choices elicited from each individual (N), (b) the method used to design the decision set (designed vs. random), (c) the outcome domain (gains only vs. mixed), (d) the nominal outcome scale, and (e) the realized incentive scale. Table 1 provides a schematic outline of the differences between these three decision sets.

Stott (2006) provided a discussion of designing his decision set suggesting that "intuitively, a good question set needs to contain a range of questions that can separate the relatively risk seeking from the relatively risk averse participants" (p. 112). Additionally, Stott's decision set was designed to minimize parameter interaction problems, where choices could be explained equally well with different combinations of parameter values. The second decision set was used by Rieskamp (2008) and was generated by Erev et al. (2002). Rieskamp defended the use of this decision set because it was generated randomly and would not give any model an unfair advantage (henceforth, Erev et al.,

Table 1
Summary of the Decision Sets

Decision set	Number of decision problems (*N*)	Generation method	Outcome domain	Absolute nominal outcome (exchange rate)	Absolute realized outcomes ($ U.S.)
Erev et al. (2002)	200	Randomly generated	Gains only	0–100 points (40:1)	$0.025–$2.50
Glöckner and Pachur (2012)	138	Mix of designed and random	Mixed	€0– €1,000 (100:1)	$0.00–$14.00
Stott (2006)	90	Experimenter designed	Gains only	£0–£40,000 (8,000:1)	$0.00–$9.25

2002, decision set). The final decision set was used by Glöckner and Pachur (2012) and consisted of a collection of decision sets from previous experiments. The authors stated that they specifically included the decision sets of Holt and Laury (2002) and Gachter, Johnson, and Herrmann (2007) because these two decision sets were designed to measure risk aversion and loss aversion. Of the three decision sets, the Glöckner and Pachur decision set is the only decision set that can estimate loss aversion because the Stott and Erev et al. decision sets are strictly in the domain of gains.

The scale of the nominal outcomes for the gambles is different for each of these decision sets. In absolute value, the Erev et al. (2002) outcomes are in the range [1 point, 100 points], the Glöckner and Pachur (2012) outcomes are in the range [€0, €1,000], and the Stott (2006) outcomes are in the range [£0, £40,000]. Stott (2006) chose to use high nominal values that were representative of an average U.K. income “to represent the important financial decisions an average person can expect to encounter in their lives” (p. 112). The remaining two studies did not provide a justification for their choice of nominal outcomes. Although not explicitly stated in these experiments, we assume that the nominal outcomes were used for parameter fitting.

Additionally, the actual payments used to incentivize choices were also different in each experiment. In absolute value, the range of realized incentives (in U.S. dollars) for each experiment was [$0.025, $2.50], [$0, $14], and [$0, $9.25] for the Erev et al., Glöckner and Pachur, and Stott decision sets, respectively. Whether we believe that participants in these experiments were making choices based on the nominal values or the realized outcomes is an important psychological question that is beyond the scope of the current paper. In general we would expect that outcome scale and the choice-sensitivity parameter are highly dependent on each other such that large nominal outcome scales with little realized value should correspond to lower choice-sensitivity estimates than small outcome scales with a lot of realized value (e.g., see Kachelmeier & Shehata, 1992).

Prior parameter distributions. We chose prior parameter distributions based on the results of previous research. The ranges for diminishing sensitivity (α), decision weighting (γ), and loss aversion (λ) have been well established in the literature (see, e.g., Glöckner & Pachur, 2012; Nilsson et al., 2011; Rieskamp, 2008). We marginalized over parameters that represent diminishing sensitivity, overweighting small probabilities, and losses looming larger than gains using the ranges in Table 2. There is less empirical evidence about appropriate ranges for choice sensitivity. Glöckner and Pachur (2012) produced median estimates of choice sensitivity in the range of $0.06 \leq \varepsilon \leq 0.13$ and Nilsson et al. (2011) produced estimates of choice sensitivity in the range of $0.18 \leq \varepsilon \leq 0.25$. Based on these results, using a low range of sensitivity with $0 < \varepsilon < 1$ should capture the population of DMs making choices in our chosen experiments. Because the units of choice sensitivity are dependent on the units of

Table 2
Parameter Ranges Used for Computing Parameter Discrimination

Parameter	Feasible range
Diminishing sensitivity: α	$0 < \alpha \leq 1$
Probability weight: γ	$0 < \gamma \leq 1$
Loss aversion: λ	$1 \leq \lambda \leq 10$
Choice sensitivity: ε	$0 < \varepsilon \leq 1$

the gamble payoffs, we first draw general comparisons of each experiment based on equated outcomes (so that the range of ε will not influence the results) and then run a simulation analysis with unequal outcomes to demonstrate the accuracy of our methodology.

We computed the expected parameter discrimination measures defined by Equations 4–7 using Monte Carlo integration implemented in MatLab. We used a uniform probability function for $p(\boldsymbol{\theta})$ to equally weight all parameter values in the ranges displayed in Table 2. This distribution provides the most conservative test of finding a lack of parameter recovery because it equally weights extreme and central parameter values. In the following analyzes, the expected parameter discrimination results are presented in two ways: (a) the total bits representing the power of the entire decision set and (b) the bits per decision problem (bits/*N*) representing the efficiency of the decision set.

Analysis with equated outcome magnitudes. We first analyzed the relative performance of each decision set when the outcome magnitudes were constrained to be equal. Outcome scale interacts with choice sensitivity complicating comparisons between decision sets. We therefore first evaluated if any of the decision sets could outperform the others when outcomes and choice sensitivity were held constant. Additionally, we explored the relationship between outcome scale and choice sensitivity.

For each decision set we computed a multiplier to rescale the maximal outcomes to be equal to 1,000. We then multiplied the equated decision sets by a common percentage that varied the maximal outcome values from 0 to 1,000. We computed for each decision set the MDP for diminishing-sensitivity, decision-weighting, and loss-aversion parameters while fixing the choice-sensitivity parameter to a constant. We computed MPD using three fixed values of choice sensitivity, 0.5, 1, and 2, representing reduced, one-to-one, and elevated sensitivity, respectively.

Figure 1 displays the MPD with varying values of outcome scales and choice sensitivity for each of the decision sets. The top row of Figure 1 shows the impact of outcome scale on total MPD in bits and the bottom row shows the impact of outcome scale on the bits per decision problem (MPD/N). These graphs show that increasing outcome magnitudes and choice sensitivity increases parameter discrimination. We

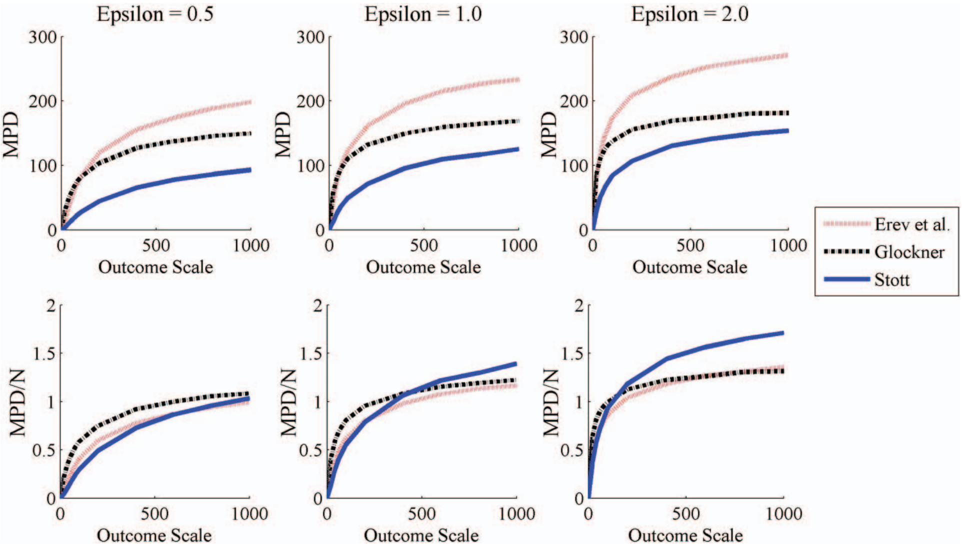

Figure 1. Expected multivariate parameter discrimination (MPD) for each decision set. The top row shows the expected discrimination achieved by the entire experiment, and the bottom row shows the expected discrimination per decision problem. Each column is based on an increasing fixed value of choice sensitivity from left to right. MPD/N = MPD bits per decision problem. See the online article for the color version of this figure.

can see that the Glöckner and Pachur (2012) decision set performs better than the other decision sets for low-outcome scales and low choice sensitivity. The efficiency of the Stott (2006) decision set increases more than the remaining decision sets with higher-outcome scales and choice sensitivity. This may be the result of the large nominal outcome scale used to design the Stott decision set. When outcome scales are equated, there is little difference between the effectiveness of these experiments on expected MPD per decision problem, despite their design differences.

This analysis shows that outcome magnitude and choice sensitivity perform similar functions of magnifying the influence of the underlying parameters in determining the subjective valuations of each gamble. Therefore, comparisons of decision sets require careful consideration of outcome scale. When outcome scale is equated, the efficiency of each of these decision sets is very similar, and the decision set with the most decision problems is predicted to produce the best results.

The relative recovery of each CPT parameter with equated outcome magnitudes. We extended the previous analysis by investigating the relative recovery of all of the free parameters using UPD and the ranges presented in Table 2. Using the same procedure as before, we equated the maximal outcome of each decision set, and varied the outcome magnitude from 0 to 1,000. Figure 2 shows the UPD as a function of maximal outcome scale for each of the CPT parameters. The results show that the diminishing-sensitivity parameter, α, has much larger parameter discrimination than the remaining parameters. On average, the expected discrimination for α is 1.5 times that of γ. Finally, the loss-aversion and choice-sensitivity parameters have much lower UPD compared to diminishing-sensitivity and decision-weighting parameters. The results in the bottom row of Figure 2, controlling for the number choices, reveal that the average expected parameter discrimination does not differ much for each of the experiments when outcome scales are equated.

Prediction of parameter recovery with unequal outcome scales. Finally, we investigated the relative performance of each of these decision sets with different outcome scales. This analysis is a theoretical exercise intended as a demonstration of the predictive ability of the methodology. The generalizability of these

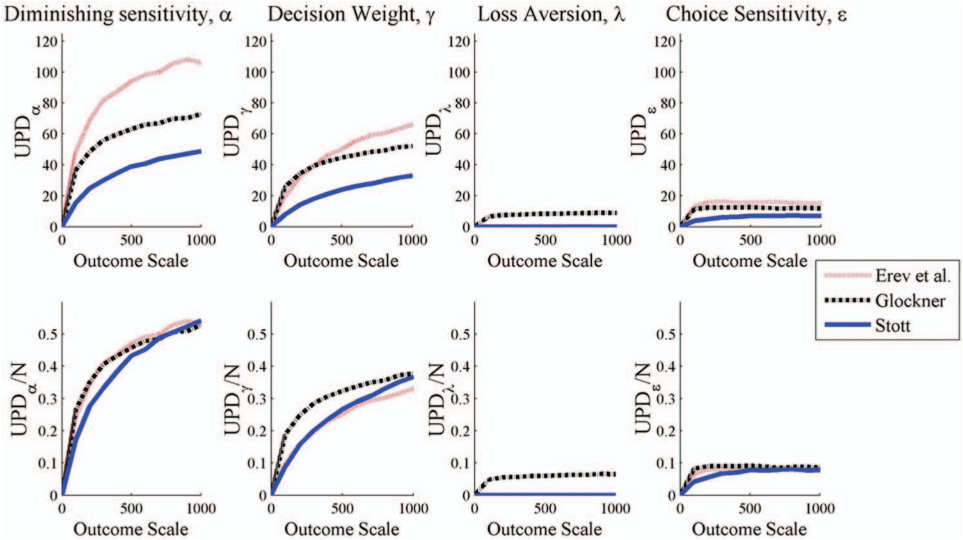

Figure 2. Expected univariate parameter discrimination (UPD) for each of the free parameters of cumulative prospect theory. The top row shows the expected discrimination achieved by the entire experiment, and the bottom row shows the expected discrimination per decision problem. UPD/N = UPD bits per decision problem. See the online article for the color version of this figure.

results is therefore limited to populations of DMs as described by Table 2. We chose to measure the usefulness of these decision sets using two different types of outcomes: *incentive compatible* and *nominal*. The incentive compatible outcome analysis was based on the realized payments of each experiment in the common currency of U.S. dollars (see Table 1). This analysis followed a purely economic perspective looking at the expected precision of parameter estimates solely as a function of the incentives used in each experiment, similar to the approach taken by von Winterfeldt and Edwards (1982). The incentive compatible outcomes are a "worst-case scenario," where DMs produced choices based on the realized payments. The nominal outcomes are a "best-case scenario," where DMs produced choices based on the larger nominal outcome scales. This analysis also included the effect-size discrimination, $\hat{d}'$, in order to evaluate parameter discrimination in both relative and absolute terms.

We present the expected parameter discrimination in total bits and bits per decision problem (bits/N) in the top row of Table 3. The left (right) hand side of the table reports the results for incentive compatible (nominal) outcomes. The predictions based on total parameter discrimination and discrimination efficiency per decision problem are the same. Our measures predict the Glöckner and Pachur (2012) decision set to be the most accurate for the lower magnitude (incentive compatible) outcomes and the Stott (2006) to be the most accurate for the higher magnitude (nominal) outcomes. Despite having more decision problems, the Erev et al. (2002) is predicted to be the least accurate decision set due to the lower outcome scales.

The second row of Table 3 displays the total bits of effect-size discrimination and the estimated equivalent effect size, $\hat{d}'$, for each parameter based on a δ of 1/10 the total range of each parameter. These results directly mimic the rank ordering produced by MPD and UPD. The effect sizes for the incentive compatible outcomes are very low ($<$1.75) for all parameters. However, the effect sizes for the nominal outcomes are much higher, with diminishing sensitivity and decision weights producing estimates greater than three. While the effect size of the diminishing sensitivity and decision weights increases substantially with higher outcome scales, the effect size of loss aversion and choice sensitivity increase more slowly. Even in the best case scenario where DMs focus on the nominal outcomes, our measures predict that loss aversion and choice sensitivity are less recoverable, replicating the predictions from the previous analysis with equated outcomes.

We additionally computed for each decision set the percent reduction in expected parameter discrimination presented in Equation 8. This measure of parameter interaction captures the reduction in expected discrimination for each parameter caused by the misspecification of the remaining parameters. The results are displayed in the third row of Table 3. In general, each decision set produced similar results across all parameters. Diminishing sensitivity has the lowest percent reduction for all three decision sets, indicating that its recovery is least affected by the values of the other parameters. The percent reduction in expected parameter discrimination was higher for the Stott (2006) decision set compared to the Erev et al. (2002) random decision set. This result is surprising because the Stott (2006) decision set was designed to reduce parameter interaction.

Simulated parameter recovery. We tested the usefulness of our measures by estimating parameters from simulated DMs with true parameter values distributed uniformly across the ranges in Table 2. We compared our measures of parameter discrimination to the root mean squared error (RMSE) of the estimated parameters from each of the three decision sets. The RMSE was chosen as a measure of error to capture both variance and bias in estimates.

The simulation was designed to match our methods for computing expected parameter discrimination by simulating choices from the CPT functional forms in Equations 12–15. The simulation ran as follows: (a) each combination of parameters were selected and taken to be the true parameter vector, (b) 100 simulated DMs made choices for the entire decision set, (c) 100 sets of parameters were estimated from these choices, and (d) the RMSE was computed between the true parameter vector and the 100 estimated parameter vectors. Finally, the RMSE of each parameter was averaged across all the parameter combinations.

We performed parameter estimation using a numerical maximum likelihood method to mini-

Table 3

Unequal Outcome Analysis of the Three Recent Decision Sets Based on (a) Expected Parameter Discrimination, (b) Effect Size Discrimination, (c) Percent Reduced Discrimination, and (d) RMSE of Simulated Parameter Recovery

	Incentive compatible outcomes						Nominal outcomes					
	Erev et al. (2002)		Glöckner and Pachur (2012)		Stott (2006)		Erev et al. (2002)		Glöckner and Pachur (2012)		Stott (2006)	
	N = 200		N = 138		N = 90		N = 200		N = 138		N = 90	
	[$0.025–$2.50]		[$0.00–$14.00]		[$0.00–$9.25]		[0–100 points]		[€0–€1,000]		[£0–£40,000]	
	Bits	Bits/N	Bits	Bits/N	Bits	Bits/N	Bits	Bits/N	Bits	Bits/N	Bits	Bits/N
Expected discrimination												
MPD	1.32	0.007	26.55	0.192	1.96	0.022	82.10	0.410	147.30	1.067	163.24	1.814
UPD_{α}	0.37	0.002	9.58	0.069	0.92	0.010	48.77	0.244	74.67	0.541	87.56	0.973
UPD_{γ}	0.41	0.002	7.47	0.054	0.53	0.006	19.55	0.098	50.13	0.363	69.46	0.772
UPD_{λ}	0.00	0.000	3.69	0.027	0.00	0.000	0.00	0.000	9.02	0.065	0.00	0.000
UPD_{ε}	0.54	0.003	5.66	0.041	0.51	0.006	13.60	0.068	12.34	0.089	5.56	0.062
	Bits	$\hat{d}'$	Bits	$\hat{d}'$	Bits	$\hat{d}'$	Bits	$\hat{d}'$	Bits	$\hat{d}'$	Bits	$\hat{d}'$
Effect size discrimination												
α with δ = 0.1	0.05	0.28	1.99	1.66	0.23	0.56	7.13	3.14	12.69	4.19	22.16	5.54
γ with δ = 0.1	0.06	0.28	0.46	0.79	0.05	0.26	2.01	1.67	6.55	3.01	13.06	4.25
λ with δ = 1.0	0.00	0.00	1.30	1.34	0.00	0.00	0.00	0.00	2.41	1.83	0.00	0.00
ε with δ = 0.1	0.11	0.38	1.61	1.50	0.16	0.47	4.28	2.44	4.71	2.55	3.75	2.28
	Percent		Percent		Percent		Percent		Percent		Percent	
Percent reduced discrimination												
PRD_{α}	0.14		0.20		0.17		0.12		0.18		0.21	
PRD_{γ}	0.15		0.23		0.22		0.23		0.23		0.25	
PRD_{λ}	—		0.20		—		—		0.31		—	
PRD_{ε}	0.17		0.23		0.27		0.24		0.26		0.42	
	Average RMSE		Average RMSE		Average RMSE		Average RMSE		Average RMSE		Average RMSE	
Simulation results												
Diminishing sensitivity: α	0.36		0.23		0.32		0.15		0.08		0.06	
Probability weight: γ	0.36		0.25		0.35		0.23		0.16		0.16	
Loss aversion: λ	—		2.74		—		—		2.09		—	
Choice sensitivity: ε	0.30		0.29		0.29		0.29		0.23		0.29	

Note. RMSE = root mean squared error.

mize $G^2 = -2 * \ln[L(\alpha, \gamma, \lambda, \varepsilon \mid x_1, \ldots, x_N)]$ (see Glöckner & Pachur, 2012; Rieskamp, 2008). Minimization was achieved by first using a grid search of the parameter values to produce an initial parameter vector for numerical minimization in Matlab using *fminsearch.m*. Due to the computational intensity of the estimation routine, true parameters were taken as the factorial combination of the sets {0.1, 0.3, 0.5, 0.7, 0.9} for the range of [0, 1] and {1, 3, 5, 7, 9} for the range [1, 10].

The simulation did not place any univariate restrictions on the parameter estimation routine; all parameters were estimated simultaneously from a single collection of *N* choices. The estimated parameters were, however, restricted such that estimates of α, γ, and ε were in the interval [.1, 1] and estimates of λ were in the interval [1, 10].[5]

The bottom row of Table 3 shows the average RMSE for the recovered parameters. As predicted by our measures, the incentive compatible experiments are associated with a large amount of noise for recovering all of these parameters when assuming a population of DMs defined by Table 2. The low range of choice sensitivity may be responsible for the poor parameter discrimination based on the realized incentives. The nominal outcome experiments are associated with improved recovery, mostly for the diminishing-sensitivity and decision-weighting parameters. Even with the much larger nominal outcome scales, estimates of choice sensitivity and loss aversion are still associated with large amounts of error.

There is a strong negative association between the normalized RMSE values (RMSE/range) and the corresponding values of UPD, showing that our measures of parameter discrimination are a good indicator of parameter recoverability for the incentive compatible outcomes (Pearson $r = -0.83$; Kendall $\tau = -0.78$), the nominal outcomes ($r = -0.93$; $\tau = -0.70$), and across both outcomes scales ($r = -0.91$; $\tau = -0.82$). Due to the differences in outcome scale, the relative performance of recovering each of the parameters varies for each decision set. The results support our previous analysis that the larger outcomes increase the recoverability of diminishing sensitivity and decision weighting at a higher rate than the recoverability of loss aversion and choice sensitivity. While we have demonstrated the predictive accuracy of our methodology, these results are limited to our choices of outcome magnitude and range of choice sensitivity.

The Description–Experience Gap

The D–E gap is an empirical finding based on differences between DMs who choose between described gambles and DMs who choose between an identical set of experienced gambles (Hertwig et al., 2004; Weber et al., 2004). Description-based choices involve making decisions between the descriptions of two gambles, similar to the gamble choices presented by Kahneman and Tversky (1979). Experience-based choices involve making decisions without descriptions, but instead using outcomes drawn randomly from a gamble (experienced samples). Research suggests that description-based choices are better predicted by an inverse "S" shaped weighting function that overweights small probabilities, whereas experience-based choices are better predicted by an "S" shaped weighting function that underweights small probabilities. This difference is known as the D–E gap.

The empirical evidence for the D–E gap consists of preference reversals between described and experienced versions of gamble pairs (Hertwig et al., 2004), with the D–E gap narrowing with more experience (Hau et al., 2008) and CPT predictions underweighting small probabilities being more predictive than CPT predictions overweighting small probabilities (Ungemach et al., 2009). In terms of decision modeling, these predictions imply different values for the decision-weighting parameter, γ. However, it is not clear how changing the representation of each gamble affects parameter recoverability. Rakow, Demes, and Newell (2008) demonstrated that presenting identical information for both descriptions and experiences reduced the D–E gap.

Several studies demonstrated the reversal of CPT predictions by using choice data (Hau et al., 2008; Hertwig et al., 2004; Ungemach et al., 2009). In general, fitting CPT parameters to experience-based choices is performed using a

[5] These restrictions are similar to those used for parameter estimation by previous research (see, e.g., Glöckner & Pachur, 2012; Nilsson et al., 2011; Rieskamp, 2008).

two-step process (Fox & Hadar, 2006): (a) the gamble outcomes and probabilities are represented in the model with the relative frequencies of each outcome that occurred during experience and (b) the CPT value and weighting functions are then applied to these outcomes and relative frequencies to produce choice predictions. Parameter estimation is typically performed by finding a set of parameters to either maximize percent correct prediction or maximize likelihood. We adopted this two-step process to measure expected parameter discrimination for experienced gambles represented by the relative frequencies of each experienced outcome.

Assuming that the relative frequency of each outcome is representative of the DM's probability estimate for each outcome, then the granularity of their probability estimates is determined by the number of draws used to produce the estimate. For example, three draws can produce relative frequencies in increments of 1/3 (e.g., 0, 1/3, 2/3, and 1), while five draws can produce relative frequencies in increments of 1/5. Therefore, a probability of 0.1 cannot be accurately estimated using relative frequencies of five draws from a binomial distribution.

We predicted that the granularity reduction in probability estimation would reduce the expected parameter discrimination corresponding to the weighting-function parameter, γ. This prediction was based on the *amplification effect* (Hertwig & Pleskac, 2008, 2010) that experiences with few draws amplify the perceived absolute mean difference between gambles, making choices seem easier. We based our prediction on the idea that the granularity reduction in probability estimates produces easier choices (i.e., larger perceived differences between gambles pairs), causing most DMs to make the same choice despite having different decision-weighting parameters. Therefore, estimates of decision weights from experience-based choices are less reliable than estimates of decision weights from the equivalent description-based choices. We also predicted that the higher estimation variance associated with experienced gambles produced a systematic bias in the estimates of decision weights, accounting for the systematic D–E gap empirical results.

Analysis of D–E gap decision problems. We analyzed the difference in parameter recovery for described versus experienced decision problems using expected parameter discrimination. We used the parameter ranges presented in Table 2 with the following exceptions. We restricted the range of true values for γ to [0.50, 1.00] and the range of alternative values for γ to [1.00, 2.00].[6] These ranges directly test the decision set on differentiating choices that overweight small probabilities ($\gamma < 1$) from choices that underweight small probabilities ($\gamma > 1$). We compiled a decision set from the literature demonstrating the D–E gap by taking the union of the decision problems used by Hertwig et al. (2004); Erev et al. (2010); Rakow et al. (2008), and Weber et al. (2004) and removing any repeated gambles. See online Supplemental Materials for a complete list of decision problems.

The parameter discrimination for the described version of the gambles was computed to get a baseline for expected parameter discrimination. To simulate the impact of experience, we ran a computer simulation that produced a sample of outcomes for each gamble using three sample sizes of 5, 10, or 15 draws per gamble. Statistical descriptions of the samples were computed based on the observed outcomes and relative frequencies. These observed relative frequencies were used as the problem sets for the experience-based version of the gamble. For example, the gambles (4, 0.2; 0, 0.8) and (3, 1) could produce the five draws of {0, 0, 0, 4, 4} and {3, 3, 3, 3, 3}, respectively. The simulation would then compute the relative frequencies of the outcomes and use the decision problem (4, 2/5; 0, 3/5) and (3, 5/5) to compute the parameter discrimination for the five draw version of these gambles. Because each set of draws is randomly generated and can differ, it is important to use averages to get a stable result. We report the average parameter discrimination across 100 replications in our analyses.

Table 4 presents the expected MPD and UPD of the D–E gap decision set for the descriptions of the gambles and for each of the sample sizes of 5, 10, and 15 draws experienced per gamble. The results in the description column show that the overall expected parameter discrimination is much higher for estimating the weighting-function parameter, γ, than for estimating the remaining parameters. This reflects the fact that

[6] The bounds for these decision sets were selected to be symmetric across the reciprocal of γ.

Table 4
The Impact of Experience on the Measures of Expected Parameter Discrimination for the Gambles Used to Investigate the Description–Experience Gap

		Experienced draws per gamble		
	Description	5	10	15
Description–experience gap set, $N = 75$				
MPD	11.72	10.08	10.71	10.91
UPD_α	2.63	4.61	3.83	3.40
UPD_γ	**6.28**	**1.30**	**3.22**	**4.07**
UPD_λ	0.78	1.20	1.11	1.04
UPD_ε	1.80	2.89	2.45	2.26

Note. MPD = multivariate parameter discrimination; UPD = univariate parameter discrimination. Boldface depicts the parameter discrimination for the weighting function parameter.

the parameter ranges for computing the expected discrimination of γ were designed to compare overweighting to underweighting. Compared to the described version of the gambles, the expected UPD for the weighting-function parameter reduces by 80% for choices based on gambles with five experienced draws per gamble. This reduction in discrimination does not occur for any of the remaining parameters. The UPD for γ increases with more experience to roughly 65% of the described version of the gambles for 15 experienced draws per gamble.

Bias in estimates of decision weights. The empirical results for the D–E gap show that decision weights are systematically higher for experienced gambles than for described gambles. We show that this result can be explained by the fact that increased estimation variance can produce systematic biases in estimates of the decision weights for DMs who actually overweight small probabilities. This is due to the fact that (a) the decision weight parameter is bounded at zero and (b) overweighting small probabilities places the theorized true parameter value close to the lower bound. A property of any positively skewed random variable bounded to be greater than zero is that as variance increases, the mean increases (e.g., the χ^2 distribution). Therefore, we predict systematic biases in estimation will reproduce the empirical results of D–E gap using simulated DMs that consistently overweight small probabilities. Specifically, we predict that relative to description-based choices, the experienced-based choices will lead to recovered parameters that are biased toward underweighting small probabilities.

We demonstrated this effect using simulated DMs set to overweight small probabilities. We estimated 1,000 decision-weighting parameters, γ, from the described and experienced version of the D–E gap decision set using simulated choices from 1,000 DMs with the parameter set $[\alpha, \gamma, \lambda, \varepsilon] = [0.5, 0.5, 1.0, 1.0]$. We selected this particular vector of parameters to have DMs who represent the findings from description-based research: diminishing sensitivity and overweighting small probabilities. Using the numerical maximum likelihood method described above, only the decision-weighting parameter was estimated such that $\gamma \in [0.1, 2]$, and the remaining parameters were set to their true value. Choices made on described gambles were fit using the described gambles and choices made on experienced gambles were fit using the exact same experienced representation. As before, the experienced versions of the gambles were based on 5, 10, and 15 draws from each gamble (see the online Supplemental Materials for a replication with alternate parameter values).

Each distribution of decision-weighting estimates is displayed in Figure 3. The true parameter value is marked at 0.5 with a dashed line. The parameter estimates from the experienced gambles are clearly more biased toward higher decision weights and have more variance than the parameter estimates from the described ver-

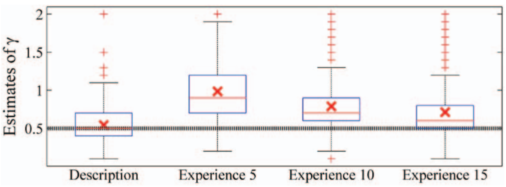

Figure 3. Distribution of decision weight estimates, γ, produced by the described and experienced version of the description–experience gap decision set. The true decision weight, $\gamma = 0.5$, is marked with a dashed line. The central 50% of the distribution of estimates is contained in the box, the whiskers cover 1.5 times the interquartile range, and outliers are marked with a plus sign. The median is marked with a line and the mean is marked with an "x." See the online article for the color version of this figure.

sion of the gambles. The results are due to the minimal information gained by the experience-based experiments to recover these parameters, allowing estimates to vary across the full range from [0.1, 2]. While the estimation bias (and variance) reduces as the number of draws increases from 5 to 15, the distribution medians never fully converge to the true value of $\gamma = 0.5$.

Summary. Our analysis of parameters recovered from experienced gambles supports the predictions that estimation precision is reduced for experience-based decision problems and that the expected parameter discrimination for estimating γ is strongly reduced when gambles are represented by samples of 5–10 draws. The decision sets designed to study the D–E gap reveal an alarming reduction in the information for estimating the weighting-function parameter, γ, relative to the diminishing-sensitivity parameter, α. Finally, we demonstrate how reduced parameter discrimination for experienced gambles can potentially generate systematically biased estimates of decision weights.

General Discussion

We developed and demonstrated the benefit of using expected parameter discrimination to assess the overall quality of a decision set for estimating parameters for CPT, currently the most widely used descriptive model of choice. This methodology is closely linked with (adaptive) design optimization, a technique developed to effectively select stimuli to accurately recover parameter estimates. Our proposed methodology answers fundamental questions in the literature of static experiments by allowing researchers to assess (a) the overall effectiveness of previous decisions sets for parameter recovery on any decision model, (b) the differential accuracy of estimating each parameter, and (c) how changes in representations of decision problems can impact parameter recovery.

We have also provided tools for easily computing and interpreting these measures in an effort to make more efficient designs easily accessible to all behavioral researchers. We used numerical methods that do not require analytical solutions to apply our measures of parameter recovery to CPT to ensure that our approach can be easily generalized to many models. Our demonstration covers many mathematical complexities including rank dependent decision weights, the separation of gains and losses, nonlinear functions, and multiple free parameters. Our measures of parameter recovery can therefore be applied to a large variety of cognitive models; the only requirement for measuring parameter recovery with our approach is a probabilistic choice prediction and a well-defined likelihood function. For this reason, our approach also extends easily to models that make predictions about continuous responses (this would require only a minor modification to the likelihood function). We conclude by discussing the novel results obtained by applying our measures to CPT and implications for future research using choice data.

The Effectiveness of Choice Data Estimation

The most important contribution of this method is to provide a framework for analyzing the viability of using choice data estimation to recover the free parameters of decision models. Considering the discriminability of parameters is as necessary as considering effect sizes (or power analyzes) in psychological research. Assuming that DM's focus on the realized incentives for each gamble, our results suggest that choice data estimation may not currently provide very accurate results. When assuming DM's focus on the much higher nominal outcome scales, the performance of each decision set improves but is still questionable. More general conclusions about the effectiveness of choice data estimation will require detailed assumptions about appropriate ranges of choice sensitivity and to determine how DMs perceive nominal versus realized outcomes.

While previous research has not completely ignored the impact of a decision set on the quality of experimental results, we provide the first formal analysis designed to explicitly quantify such differences. For example, when gamble outcome scales were equated, we found little difference between the performance of a randomly generated and experimenter designed decision set. Each performed with roughly equal efficiency per decision problem. More surprisingly, the Stott (2006) decision set was designed to reduce parameter interaction but our measures indicate more parameter interaction in the Stott decision set than in the randomly generated Erev et al. (2002) decision set. Designing decision sets to minimize parameter interactions is a difficult task that can be more easily achieved with our proposed measures. These results provide strong evidence for the value of our metrics for designing future studies.

Parameters Differ in Their Recoverability

The relative recoverability of CPT parameters has received little normative attention. When CPT is assumed to be the true decision model, our results based on equated outcomes show stark differences in parameter recovery across a large range of outcome scales (see Figure 2). Our general findings are that the easiest parameters to estimate are the diminishing-sensitivity and decision-weighting parameters. The hardest parameters to estimate are the loss-aversion and choice-sensitivity parameters, which achieve roughly half the information compared to diminishing sensitivity.

With regard to the loss aversion, the Glöckner and Pachur (2012) decision set was the only experiment specifically designed to estimate this parameter. Based on the poor recoverability of the loss-aversion parameter, we do not believe it is coincidence that several studies debate the existence of loss aversion all together (see Ert & Erev, 2013, for a review). Because loss aversion cannot be estimated without negative outcomes, only choices involving mixed gambles can contribute toward estimating this parameter. For example, half the decision problems in the Glöckner and Pachur decision set had negative outcomes and roughly a quarter of the decision problems involved mixed gambles. However, as demonstrated by our analysis of the D–E gap, it can be more difficult to identify decision problems that fail to contribute toward estimating the remaining parameters, especially decision weighting. Future tests of parameters, such as loss aversion, can be performed under more controlled circumstances by using our measure of expected parameter discrimination to equate static decisions sets in their ability to estimate any desired parameter.

Our simulation analysis demonstrates the complications of predicting relative performance of decision sets when outcomes scales are not equated. The experiments with lower payments were associated with better recovery of choice sensitivity because the randomness in the choices reduced the ability to recover the remaining parameters. The experiments with higher payments produced choices that were less random, and the other parameters became easier to recover. Overall, the small incentives used in each experiment generated large errors for recovering all of the parameters. Therefore, results of such experiments may be less stable if the DMs fail to focus on the nominal payments and their choice sensitivity is less than one. Future research should further investigate the tradeoff between outcome scale and choice sensitivity. For example, is it more efficient to use larger nominal outcomes that will be associated with lower choice sensitivity or larger realized payments that will be associated with higher choice sensitivity?

While we can investigate the value of a static decision set for recovering parameters, the reverse analysis can also be informative. If a particular parameter cannot be reliably estimated from particular classes of stimuli, we can conclude that many different values of this parameter generate the same choices for these stimuli, and that specific values of this parameter are not as important for the model's predictions. This type of conclusion is precisely the question one needs to ask when evaluating the importance of decision weights for making choices between experienced gambles.

Empirical Evidence for the Description–Experience Gap

There has been much research on the D–E gap seeking to examine the impact of experience on the shape of the weighting function. We find that experience can drastically reduce the information available for differentiating a DM's true decision-weighting parameter from an alternative parameter. In accordance with the D–E gap empirical evidence mentioned above, our analysis shows that increasing the number of experienced draws increases the parameter recoverability. Moreover, our analysis shows that few experienced draws change the nature of the choices that DM's are making, similar to the amplification effect (Hertwig & Pleskac, 2008, 2010), and reduces the impact of different types of decision weights to produce different choices. These results support a similar analysis from Broomell, Budescu, and Por (2011) that found high levels of overlap between a variety of models attempting to predict choice from experiences. Our results also complement results from Rakow et al. (2008) where the D–E gap is reduced when DMs see the exact same summaries of decision problems. Finally, relative to decision weights estimated from described gambles, decision weights estimated from experienced gambles can lead to systematically biased estimates that correspond with the empirical results of the D–E gap.

There are other psychological processes that can be linked to the continued D–E gap in very large sample sizes where relative frequencies converge to the described versions of the gambles. Future research could easily incorporate theories of number bias (Baird & Noma, 1978, p. 109) to show how subjective estimates of probability that are rounded to numbers typical of subjective judgment (e.g., 0.01, .05, .25, etc.) influence the expected parameter discrimination for estimating decision weights from choice.

Implications for Future Research

Our framework can also be used as a method for calibrating static experiments for specific research goals. Improved decision sets can be created with the goal of achieving a minimal standard of expected parameter discrimination for estimating a set of parameters from a theorized distribution of participants. One method for creating improved decision sets involves applying our measures to the individual decision problems. Decision sets can be created by ordering a collection of decision problems based on estimation precision and retaining the subset of decision problems that meet a minimal standard. Additionally, researchers seeking optimal designs should apply the Bayes D-optimum criterion to maximize the joint information gained by a decision set or, better yet, employ adaptive design optimization.

Static experiments must be designed to match the statistical assumptions of the estimation procedure. One important assumption of currently used maximum likelihood routines involves treating each choice as an independent observation. In our framework, we carried this assumption forward to solve for expected parameter discrimination of an entire decision set as the sum of the expected parameter discrimination from each decision problem. Therefore, violations of the independence assumption can easily lead to poorly designed decision sets when using expected parameter discrimination. Consider the extreme case where a single decision problem produces more information gain than any other decision problem. The maximal expected parameter discrimination for a decision set of size *N* is achieved with *N* repetitions of the same decision problem. However, the actual parameter discrimination of the entire experiment will collapse if, as is likely, the DM remembers previous choices and violates the assumption of choice independence. This problem is not unique to our measures and will have the same deleterious effect on any estimation routine.

There are many estimation methods that could be used besides maximum likelihood that also deserve consideration, such as Bayesian hierarchical estimation (Nilsson et al., 2011) and the use of proper scoring rules (Merkle & Steyvers, 2013; Selten, 1998). Some of these approaches are especially desirable as they control for the effect of prediction extremity in model comparisons. Likelihood-based approaches, in contrast, favor models that make moderate probabilistic predictions over extreme probabilistic predictions. Future work should consider extending the proposed approach to a variety of estimation methods.

References

Baird, J. C., & Noma, E. (1978). *Fundamentals of Scaling and Psychophysics*. New York, NY: Wiley and Sons.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115,* 463–501.

Broomell, S. B., Budescu, D. V., & Por, H. H. (2011). Pair-wise comparisons of multiple models. *Judgment and Decision Making, 6,* 821–831.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological review, 100,* 432–459.

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., & Pitt, M. A. (2012). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management Science, 59,* 358–375. doi:10.1287/mnsc.1120.1558

Cavagnaro, D. R., Pitt, M. A., & Myung, J. I. (2011). Model discrimination through adaptive experimentation. *Psychonomic Bulletin Review, 18,* 204–210.

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York, NY: Wiley.

Donkers, B., Melenberg, B., & Van Soest, A. (2001). Estimating risk attitudes using lotteries: A large sample approach. *The Journal of Risk and Uncertainty, 22,* 165–195.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., . . . Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making, 23,* 15–47.

Erev, I., Glozman, I., & Hertwig, R. (2008). What impacts the impact of rare events. *Journal of Risk and Uncertainty, 36,* 153–177.

Erev, I., Roth, A. E., Slonim, R. L., & Barron, G. (2002). Combining a theoretical prediction with experimental evidence. Retrieved from http://ssrn.com/abstract=1111712. doi:10.2139/ssrn.1111712

Ert, E., & Erev, I. (2013). On the descriptive value of loss aversion in decisions under risk: Six clarifications. *Judgment and Decision Making, 8,* 214–235.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114,* 817–868.

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American Psychologist, 46,* 835–847.

Fischhoff, B. (2005). Cognitive processes in stated preference methods. In K.-G. Maler & J. Vincent (Eds.), *Handbook of environmental economics* (pp. 937–968). Amsterdam, The Netherlands: Elsevier.

Fox, C., & Hadar, L. (2006). Decisions from experience = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber, & Erev (2004). *Judgment and Decision Making, 1,* 159–161.

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

Gachter, S., Johnson, E., & Herrmann, A. (2007). Individual-level loss aversion in riskless and risky choices. Retrieved from http://ssrn.com/abstract=1010597

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

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

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychphysics*. New York, NY: Wiley.

Harrison, G. W., Humphrey, S. J., & Verschoor, A. (2009). Choice under uncertainty: Evidence from Ethiopia, India and Uganda. *The Economic Journal, 120,* 80–104.

Hau, R., Pleskac, T. J., Kiefer, J., & Hertwig, R. (2008). The description–experience gap in risky choice. The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making, 21,* 493–518.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539.

Hertwig, R., & Pleskac, T. J. (2008). How small samples render choice simpler. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 209–235). Oxford, England: Oxford University Press.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition, 115,* 225–237.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review, 92,* 1644–1655.

Kachelmeier, S. J., & Shehata, M. (1992). Examining risk preferences under high monetary incentives: Experimental evidence from the People's Republic of China. *The American Economic Review, 82,* 1120–1141.

Kahneman, D., & Tversky, A. (1979). Prospect theory. *Econometrica, 47,* 263–291.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Statistical Methodology, 21,* 272–319.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22,* 79–86.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89,* 46–55.

Luce, R. D. (1992). A theory of certainty equivalents for uncertain alternatives. *Journal of Behavioral Decision Making, 5,* 201–216.

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10,* 292–304.

Myung, J. I., Cavagnaro, D. R., & Pitt, M. A. (2013). A tutorial on adaptive design optimization. *Journal of mathematical psychology, 57,* 53–67.

Nilsson, H., Rieskamp, J., & Wagenmakers, E. (2011). Hierarchical Bayesian parameter estimation for cumulative prospect theory. *Journal of Mathematical Psychology, 55,* 84–93.

Pachur, T., Hanoch, Y., & Gummerum, M. (2010). Prospects behind bars: Analyzing decisions under risk in a prison population. *Psychonomic Bulletin and Review, 17,* 630–636.

Por, H. H., & Budescu, D. V. (2013). Revisiting the gain–loss separability assumption in prospect theory. *Journal of Behavioral Decision Making, 26,* 385–396. doi:10.1002/bdm.1765

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

Rakow, T., Demes, K. A., & Newell, B. R. (2008). Biased samples not mode of presentation: Reexamining the apparent underweighting of rare events in experience-based choice. *Organizational Behavior and Human Decision Processes, 106,* 168–179.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1146–1465.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics, 1,* 43–62.

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32,* 101–130.

Tversky, A., & Fox, C. R. (1995). Weighting risk and uncertainty. *Psychological Review, 102,* 269–283.

Tversky, A., & Kahneman, D. (1992). Cumulative prospect theory: An analysis of decision under uncertainty. *Journal of Risk and Uncertainty, 5,* 297–323.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversal. *The American Economic Review, 80,* 204–217.

Tversky, A., & Thaler, R. H. (1990). Preference reversals. *The Journal of Economic Perspectives, 4,* 201–211.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted, when rare outcomes are experienced (rarely)? *Psychological Science, 20,* 473–479.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67,* 575–588.

Von Winterfeldt, D., & Edwards, W. (1982). Costs and payoffs in perceptual research. *Psychological Bulletin, 91,* 609–622.

Wang, C., & Chang, H. (2011). Item selection in multidimensionsal computerized adaptive testing-gaining information from different angles. *Psychometrika, 76,* 363–384.

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111,* 430–445.

Wu, G., & Markle, A. B. (2008). An empirical test of gain-loss separability in prospect theory. *Management Science, 54,* 1322–1335.

Appendix A

Proof of Equation 5

Proposition

The measure of expected MPD in Equation 4 is an upper bound for the Bayes D-optimum criterion for optimal designs.

$$\begin{aligned}\text{Bayes D-optimum} &= E_{\mathbf{x}}[D_{KL}(p(\boldsymbol{\theta} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{\theta}))] = I(\boldsymbol{x};\boldsymbol{\theta}) \\ &= E_{p(\boldsymbol{\theta}_0)}\left[D_{KL}(p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) \,\|\, p(\boldsymbol{x}))\right] \leq E_{p(\boldsymbol{\theta}_0)}E_{p(\boldsymbol{\theta}_1)}\left[D_{KL}(p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) \,\|\, p(\boldsymbol{x} \mid \boldsymbol{\theta}_1))\right].\end{aligned} \tag{18}$$

Proof

We can solve for the relationship between these two criteria using Jensen's inequality $(-E[\log(x)] \geq -\log[E(x)])$.

$$\begin{aligned}\text{Bayes D-optimum} &= E_{p(\boldsymbol{\theta}_0)}\left[D_{KL}(p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) \,\|\, p(\boldsymbol{x}))\right] = E_{p(\boldsymbol{\theta}_0)}\left[D_{KL}(p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) \,\Big\|\, E_{p(\boldsymbol{\theta}_1)}[p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)])\right] \\ &= E_{p(\boldsymbol{\theta}_0)}\left[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}\log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}{E_{p(\boldsymbol{\theta}_1)}\left[p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)\right]}\right] \\ &= E_{p(\boldsymbol{\theta}_0)}[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) - \log E_{p(\boldsymbol{\theta}_1)}[p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)]]] \\ &\leq E_{p(\boldsymbol{\theta}_0)}[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) - E_{p(\boldsymbol{\theta}_1)}[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)]]] \\ &= E_{p(\boldsymbol{\theta}_0)}[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}[E_{p(\boldsymbol{\theta}_1)}[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) - \log p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)]]] \\ &= E_{p(\boldsymbol{\theta}_0)}[E_{p(\boldsymbol{\theta}_1)}[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}[\log p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) - \log p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)]]] \\ &= E_{p(\boldsymbol{\theta}_0)}\left[E_{p(\boldsymbol{\theta}_1)}\left[E_{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}\log \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}_0)}{p(\boldsymbol{x} \mid \boldsymbol{\theta}_1)}\right]\right] \\ &= E_{p(\boldsymbol{\theta}_0)}E_{p(\boldsymbol{\theta}_1)}[D_{KL}(p(\boldsymbol{x} \mid \boldsymbol{\theta}_0) \,\|\, p(\boldsymbol{x} \mid \boldsymbol{\theta}_1))]\end{aligned}$$

This proof shows that the expected parameter discrimination is an upper bound on the information gain of the experiment. That means that experiments with very low parameter discrimination are constricted in the potential information that can be gained by the experiment. While the maximum of these two measures is not identical, experiments with larger parameter discrimination will perform better, on average, in discriminating the parameters of a randomly drawn DM from another randomly drawn DM.

(*Appendices continue*)

Appendix B

Proof of Equation 9

Proposition

Let $p(x) = N(\mu_0,1)$ and $q(x) = N(\mu_0 + d', 1)$. The KL divergence between these distributions $KL_{bits} = K(p(x) \parallel q(x))$ is a function of *d'*, given by the following equality:

$$KL_{bits} = \frac{(d')^2}{2 * \ln(2)}$$

Proof

Let $p(x) = N(\mu_0,1)$ and $q(x) = N(\mu_0 + d',1)$. Compute the KL divergence using natural logarithms to get KL divergence in nats.

$$\begin{aligned}
K_{nats}(p(x) \parallel q(x)) &= \int_x p(x)\left[\frac{\ln(p(x))}{\ln(q(x))}\right] = \int_x p(x) * \left[\ln(p(x)) - \ln(q(x))\right] \\
&= \int_x p(x) * \left[\ln\left(\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-\mu_0)^2}{2}}\right) - \ln\left(\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-(\mu_0+d'))^2}{2}}\right)\right] \\
&= \int_x p(x) * \left[\ln\left(\frac{1}{\sqrt{2\pi}}\right) + \frac{-(x-\mu_0)^2}{2} - \ln\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{-(x-(\mu_0+d'))^2}{2}\right] \\
&= \int_x p(x) * \left[\frac{(x-(\mu_0+d'))^2}{2} - \frac{(x-\mu_0)^2}{2}\right] = \frac{1}{2}\int_x p(x) * [(x-(\mu_0+d'))^2 - (x-\mu_0)^2] \\
&= \frac{1}{2}\int_x p(x) * \left[2d'\mu_0 - 2d'x + (d')^2\right] \\
&= \frac{1}{2}\left[2d'\mu_0\int_x p(x) - 2d'\int_x p(x) * x + (d')^2\int_x p(x)\right] \\
&= \frac{1}{2}\left[2d'\mu_0 - 2d'\mu_0 + (d')^2\right] = \frac{(d')^2}{2}
\end{aligned}$$

Convert the KL divergence in nats to KL divergence in bits by dividing by the natural log of 2.

$$KL_{bits} = \frac{K_{nats}}{\ln(2)} = \frac{(d')^2}{2 * \ln(2)}$$

Solve for d' and the mean shift between two standard normal distributions can be computed as a function of the KL divergence (in bits) through the following equality:

$$d' = \sqrt{2 * \ln(2) * KL_{bits}}$$

Received November 12, 2013
Revision received June 20, 2014
Accepted July 15, 2014 ■