2018

Heterogeneity and Parsimony in Intertemporal Choice

Michel Regenwetter
University of Illinois at Urbana-Champaign, regenwet@illinois.edu

Daniel R. Cavagnaro
California State University, Fullerton, cavdaddy@gmail.com

Anna Popova
Dell Research Labs, anna.v.popova@gmail.com

Ying Guo
University of Illinois at Urbana-Champaign, yingguo2@illinois.edu

Chris Zwilling
University of Illinois at Urbana-Champaign, zwillin1@illinois.edu

See next page for additional authors

Regenwetter, Michel; Cavagnaro, Daniel R.; Popova, Anna; Guo, Ying; Zwilling, Chris; Lim, Shiau Hong; and Stevens, Jeffrey R., "Heterogeneity and Parsimony in Intertemporal Choice" (2018). *Jeffrey Stevens Papers & Publications*. 13.
https://digitalcommons.unl.edu/psychstevens/13

Authors
Michel Regenwetter, Daniel R. Cavagnaro, Anna Popova, Ying Guo, Chris Zwilling, Shiau Hong Lim, and Jeffrey R. Stevens

Heterogeneity and Parsimony in Intertemporal Choice

Michel Regenwetter
University of Illinois at Urbana-Champaign, USA, regenwet@illinois.edu

Daniel Cavagnaro
California State University at Fullerton, cavdaddy@gmail.com

Anna Popova
Dell Research Labs, USA, anna.v.popova@gmail.com

Ying Guo, Chris Zwilling
University of Illinois at Urbana-Champaign, USA, yingguo2@illinois.edu, zwillin1@illinois.edu

Shiau Hong Lim
National University Singapore, shonglim@gmail.com

Jeffrey R. Stevens
Max Planck Institute for Human Development and University of Nebraska-Lincoln, jeffrey.r.stevens@gmail.com

September 19, 2016

Corresponding Author:
Michel Regenwetter
Department of Psychology
603, E. Daniel Street, Champaign, IL 61820
regenwet@illinois.edu
www.regenwetterlab.org

Abstract

Behavioral theories of intertemporal choice involve many moving parts. Most descriptive theories model how time delays and rewards are perceived, compared, and/or combined into preferences or utilities. Most behavioral studies neglect to spell out how such constructs translate into heterogeneous observable choices. We consider several broad models of transitive intertemporal preference and combine these with several mathematically formal, yet very general, models of heterogeneity. We evaluate 20 probabilistic models of intertemporal choice using binary choice data from two large scale experiments. Our analysis documents the interplay between heterogeneity and parsimony in accounting for empirical data: We find evidence for heterogeneity across individuals and across stimulus sets that can be accommodated with transitive models of varying complexity. We do not find systematic violations of transitivity in our data. Future work should continue to tackle the complex trade-off between parsimony and heterogeneity.

Key Words: Heterogeneity; Intertemporal Choice; Noise; Random Preference; Transitivity of Preferences.

In press at *Decision*.

1 Introduction

A dieter must choose between the immediate gratification of a waistline-expanding piece of cake or the longer-term health benefits of fruit. A business manager must choose between developing projects with 'low-hanging fruit' or investing time, personnel, and money into achieving long-term goals of the firm. From diet choices to large-scale organizational decisions, all such *intertemporal choices* involve options available at different points in time (Read, 2004). In this paper, we consider binary choice between one delayed reward and another that is larger in size but also requires a longer wait. Such pairwise choices are highly heterogeneous in that they vary across decision makers and within a given decision maker over repeated decisions within even short time periods.

Economists and psychologists have developed dozens of models for intertemporal choices aimed at understanding how decision makers trade off between smaller, sooner and larger, later rewards. Most of these are *temporal discounting* models that generate a subjective present value for an option discounted by the time delay to receiving the reward. For instance, $100 in one year is less valuable than $100 in a week, which, in turn, is still less valuable than $100 today. Discounting models that map rewards and time delays to numerical subjective values of time-delayed rewards, such as exponential and hyperbolic discounting, imply transitive preferences according to which a person preferring x to y and y to z must prefer x to z (see, e.g., Doyle, 2013; Doyle and Chen, 2012; Ebert and Prelec, 2007; Frederick et al., 2002; Green and Myerson, 2004; Killeen, 2009; Laibson, 1997; Loewenstein and Prelec, 1992; Mazur, 1987; McClure et al., 2007; Samuelson, 1937).[1]

The study of the fundamental nature of intertemporal preferences faces a profound challenge. Existing tests of intertemporal choice theories rarely account explicitly for heterogeneity in behavior within and between people. It may not be possible to select a 'good' theory of intertemporal choice unless this theory jointly accounts for core preferences and heterogeneity in behavior. In our view, if we are to understand intertemporal choices, we should develop a rigorous approach that incorporates individual differences, variability in choices, and generalizability across stimuli. Therefore, rather than attend to the specifics of core preferences, such as the functional form of discounting curves, and rather than seek out a 'best' theory, we focus in this paper on the complicated interplay between parsimony and empirical variability. We also concentrate on transitive intertemporal preference and how it manifests itself in probabilistic choice. Combining transitivity of preferences with the trade-off between parsimony and variability fills a gap in the existing literature in intertemporal choice by zooming out to a broad class of theories while zooming in to the sources and types of heterogeneity.

Accounting for heterogeneity comes at the cost of reducing model parsimony. Intuitively, an excessively parsimonious model may only account for one choice made by one person at one time point for one particular stimulus. Such an overly specific model is unlikely to generalize to other stimuli presented to the same person, to other occasions on which the same person is presented with the same stimulus, to other individuals, and/or to other stimuli. At the other end of the spectrum, a model that universally accounts for the behavior of all of humanity, at all times, and over all conceivable intertemporal stimuli may have to be overly flexible. Clearly, we need to aim for some sort of middle ground. It is therefore not surprising that much of the literature in decision research, and intertemporal choice in particular, aims merely at modeling the prototypical decision maker or at documenting trends and significant effects. Though this may be useful, it could also be inherently misleading in that almost no actual person might act like that 'prototypical' decision maker. We unpack the intimate connection between models of heterogeneity in preferences and in responses for transitive theories of intertemporal preference. We also explore how adequate theoretical accounts may vary with the stimuli used. We believe that careful attention to the nature and sources of heterogeneity is essential to advancing our understanding of intertemporal choice.

Without a good theory of heterogeneity, scholars risk making too many modifications in the functional forms of core theories in an effort to accommodate "discrepancies" between theory and data, when, instead, they should model the sources of heterogeneity of behavior more explicitly. This paper provides a roadmap for accomplishing the latter by formally spelling out two major sources of heterogeneity: probabilistic responses and probabilistic preferences. We then show that these sources of heterogeneity can be incorporated into theories of intertemporal choice at an abstract level. We take a big-picture perspective and tackle intertemporal choice at a somewhat abstract level. We consider general classes of core models that share one or more of the features that 1) preferences are transitive linear orders, 2) choice options are represented by numerical utilities, 3) strengths of preferences are consistent with transitive preferences. Likewise, we consider general classes of probabilistic mechanisms for pairwise choice, namely 1) aggregation-based models that encompass various response error models as special cases and 2) distribution-free random preference, random function and random utility models that model the

[1] Other models, such as the "similarity" and "tradeoff" models, permit intransitive preferences (see, e.g., Read, 2001; Leland, 2002; Rubinstein, 2003; Scholten and Read, 2006, 2010; Stevens, 2016; Manzini and Mariotti, 2006). Here, a person may prefer x to y and y to z, yet prefer z to x for some x, y, z. A separate (companion) paper tests non-transitive heuristic models on different stimuli and different respondents.

preferences themselves as uncertain. This approach to heterogeneity is conceptually and mathematically different from the common approach that aims to accommodate individual differences through refining the core functional form of a theory, e.g., by adding extra parameters that permit specific kinds of flexibility in the core theory. Instead, our approach resembles the literature on axiom testing in decision making in that we consider the general axiom of transitivity together with general classes of probabilistic specifications.

A major strength of our approach is that it allows triage of entire classes of theories. Nonetheless, even within this general and abstract paradigm of transitivity of intertemporal preference, the number of models to consider is substantial, and different models differ dramatically in their parsimony. Furthermore, investigating the tradeoff between parsimony and heterogeneity is computationally costly. Because we consider 20 probabilistic models separately for 61 individual decision makers on six different stimulus sets, because we employ both frequentist and Bayesian analysis methods, and because many of our analyses utilize either grid search or Monte Carlo sampling methods, our analyses necessitated the use of supercomputing resources.[2]

We first discuss how to spell out a model of binary choice behavior for a person with transitive preferences. We emphasize that, in contrast to the risky choice literature, the intertemporal choice literature has largely neglected modeling the sources and types of uncertainty that underlie probabilistic behavioral data. We fill this gap by introducing eight types of probabilistic choice models of transitive intertemporal preference. After we review suitable statistical analysis methods and two experiments, we give an in-depth report on quantitative analyses at the individual and group level. We particularly highlight how parsimony trades off with accounting for within- and between-person heterogeneity. In contrast to previous such projects, we concentrate on intertemporal choice.

2 Transitive intertemporal preference and choice

In behavioral science, it is crucial not to mistake models of hypothetical constructs for models of observable behavior. The literature on intertemporal choice engages in a thorough discussion about hypothetical constructs such as preference or utility, while usually omitting a detailed model of observable behavior such as choice. We review probabilistic choice models aimed at formally representing the uncertainty that is inherent in overt behavior. We then walk through the step-by-step approach to design and test an explicitly specified theory of pairwise intertemporal choice. Since any real collection of experiments can only utilize finitely many stimuli, we assume throughout, and without much loss of generality, that the set of all choice alternatives under consideration is finite. We also concentrate on the common experimental paradigm of pairwise choice between a larger reward available with a longer delay and a smaller reward available with less delay.

2.1 Preference

Many models of binary preference between a larger, later reward L and a smaller, sooner reward S characterize a three-component cognitive process: They specify implicitly or explicitly how a decision maker 1) subjectively perceives time, 2) subjectively perceives rewards, and 3) subjectively perceives the interaction between time and rewards. This permits them to define such hypothetical constructs as the pairwise preference among choice options, the subjective value of an option, or the subjective strength of preference among pairs of options. In addition, in order to actually predict or explain behavior, a model must specify how hypothetical constructs such as subjective values or preferences translate into something one can observe, such as overt choice behavior. Before discussing choice, we start by reviewing models of transitive intertemporal preference.

A broad class of theories for intertemporal preference uses numerical functions and operations on numbers to model either subjective values of options or subjective strengths of preference among options. Suppose that x is the option of receiving a monetary or nonmonetary reward A after a time delay $t \geq 0$ (with $t = 0$ denoting an immediate reward). Many numerical models, especially many discounting models, assume that reward A is mapped into a numerical value via some value function v, that time delay t is mapped into a numerical value via some time weighting function Ψ, and that these numerical values are combined into an overall numerical value for x via some mathematical operation $\odot$, to yield an overall subjective numerical value $u(x)$ for option x as

$$u(x) = v(A) \odot \Psi(t). \tag{1}$$

[2] We ran the most computationally expensive analyses on Pittsburgh Supercomputer Center's *Blacklight* and *Greenfield* supercomputers, as an *Extreme Science and Engineering Discovery Environment* project (see also Towns et al., 2014). The analyses in this paper expended about 24,000 CPU hours on the supercomputer and more than a thousand hours on the PC.

Using this representation, many models of intertemporal preference model the preference $\succ$ as

$$L \succ S \Leftrightarrow u(L) > u(S), \tag{2}$$

where $L \succ S$ denotes that L is strictly preferred to S (see also Doyle, 2013, for similar formulations). Such a binary preference relation $\succ$ is *transitive* in that, for any options x, y, z, whenever $x \succ y$ and $y \succ z$, it follows from the right hand side of Condition 2 that $x \succ z$ as well. The general approach (1)-(2) encompasses the vast majority of theories for intertemporal choice, including the bulk of discounting models. Different implementations of such theories vary in their assumptions about the specific functional forms of v and Ψ and the operation $\odot$: Different theories use different functions $v(A)$, oftentimes focusing on quantitative rewards $A \in \mathbb{R}^+$, such as money,

$$v(A) = \begin{cases} \alpha A & \text{(often with } \alpha = 1\text{, Samuelson, 1937; Mazur, 1984),} \\ A^\alpha & \text{(Killeen, 2009),} \\ \ldots, & \end{cases} \tag{3}$$

different functions $\Psi(t)$,

$$\Psi(t) = \begin{cases} \delta^t & \text{(Samuelson, 1937),} \\ \delta t^\beta & \text{(Killeen, 2009),} \\ \frac{1}{1+\delta t} & \text{(Mazur, 1984),} \\ \frac{1}{1+\delta t^\beta} & \text{(Mazur, 1987),} \\ \frac{1}{(1+\delta t)^{\beta/\delta}} & \text{(Loewenstein and Prelec, 1992; Green and Myerson, 2004),} \\ e^{-(\delta t)^\beta} & \text{(Ebert and Prelec, 2007),} \\ \omega e^{-\delta t} + (1-\omega)e^{-\beta t} & \text{(McClure et al., 2007),} \\ \ldots, & \end{cases} \tag{4}$$

and different operations $\odot$,

$$v(A) \odot \Psi(t) = \begin{cases} v(A) \times \Psi(t) & \text{(Samuelson, 1937; Laibson, 1997; Mazur, 1984),} \\ v(A) - \Psi(t) & \text{(Killeen, 2009; Doyle and Chen, 2012),} \\ \ldots\ . & \end{cases} \tag{5}$$

(The cited papers also provide permissible ranges for the parameters $\alpha, \beta, \delta, \omega$ in these functions.)

Even the two examples of v in Eq. 3, seven examples of Ψ in Eq. 4, and two operators $\odot$ in Eq. 5 permit $2 * 7 * 2 = 28$ different combinations. The intertemporal choice literature has generated a panoply of such models for preferences, subjective values, or strengths of preferences. Most studies stop with the derivation of these constructs and do not specify response mechanisms that convert hypothetical constructs into predictions about heterogeneous overt choice behavior. Some scholars have recently started to incorporate stochastic specifications of response processes into theories of intertemporal choice (Arfer and Luhmann, 2015; Dai and Busemeyer, 2014; Ericson et al., 2015).

The fact that most theories of intertemporal choice are silent about the response mechanism is problematic. Scholars in other domains, most notably in risky choice, have warned not to think of response mechanisms as a mere optional add-on that one selects based on convenience or subjective taste of what constitutes an elegant model (Carbone and Hey, 2000; Hey and Orme, 1994; Hey, 2005; Loomes and Sugden, 1995; Loomes et al., 2002; Luce, 1959, 1995; Luce and Narens, 1994; Luce and Suppes, 1965; McCausland and Marley, 2014). Misspecification of response processes substantially affects conclusions about parameter values and readily distorts the functional form of the underlying core algebraic model (Blavatskyy and Pogrebna, 2010; Stott, 2006; Wilcox, 2008). Mis- and over-specification also compromise one's ability to predict future choices based on best-fitting parameter values in a current study. An additional formidable challenge, compounded with the suitable selection of response models, often lies in finding suitable statistical methods (Iverson and Falmagne, 1985; Myung et al., 2005; Davis-Stober, 2009). Our models and methods tackle these challenges at a high level of generality. Rather than look for a 'best' model, we focus on the interplay between heterogeneity and parsimony.

2.2 Preference and choice

We now review major model classes of probabilistic choice. We assume throughout the rest of the paper that there are only finitely many choice options under consideration, hence we always only consider finitely many binary choice probabilities.

Tremble models build on the hypothetical construct of binary preference. They start from the premise that the decision maker has a fixed "true" preference $\succ$, and that choice probabilities reflect a tendency to make occasional errors in revealing the underlying hypothetical construct. In a tremble model, it is usually assumed that the error rate for a given pair of options (x, y) is a free parameter ϵ_{xy} (Birnbaum, 2008; Birnbaum and Navarrete, 1998; Harless and Camerer, 1994), so that the probability P_{xy} of choosing x over y is

$$P_{xy} = \begin{cases} 1 - \epsilon_{xy} & \text{if } x \succ y, \\ \epsilon_{xy} & \text{if } y \succ x, \end{cases} \qquad \text{with, usually, } 0 < \epsilon_{xy} \leq \frac{1}{2}.$$

Similarly, *Fechnerian models* are based on the notion that a decision maker has a fixed "true" utility function, but because of random noise, the decision maker reveals the underlying hypothetical construct only probabilistically. In contrast to tremble models, Fechnerian models explicitly model error rates as a monotonically decreasing function of the strength of preference, $\mathbb{S}_{xy}$, with choices for strongly preferred options (large values of $|\mathbb{S}_{xy}|$) being close to deterministic and choices for extremely weakly preferred options (small values of $|\mathbb{S}_{xy}|$) resembling the toss of a fair coin (Hey and Orme, 1994; Manski and McFadden, 1981; McFadden, 2001; Thurstone, 1927). According to a Fechnerian model, the binary choice probability is given by

$$P_{xy} = F\big(\mathbb{S}_{xy}\big), \qquad \text{with } F \text{ a cumulative distribution function and } F(0) = \frac{1}{2}.$$

A logistic cumulative distribution function (CDF) yields the well-known *logit* model and a normal CDF yields the *probit* model, respectively.[3]

The strength of preference $\mathbb{S}_{xy}$, in turn, is another hypothetical construct, often derived from u using another operation, $\ominus$, via $\mathbb{S}_{xy} = u(x) \ominus u(y)$. Examples include $\mathbb{S}_{xy} = u(x) - u(y)$ or, for $u > 0$, $\mathbb{S}_{xy} = ln\left(\frac{u(x)}{u(y)}\right)$. The latter is used in a historically prominent Fechnerian model called *Luce's Choice Axiom* (Luce, 1959; Yellott, 1977), together with a unit-scaled logistic CDF, $F(x) = \frac{1}{1+e^{-x}}$, giving

$$P_{xy} = \frac{u(x)}{u(x) + u(y)}, \qquad \text{with } u(x), u(y) > 0.$$

These two response models, tremble and Fechner, treat the decision maker's hypothetical constructs (preference, utility, strength of preference) as deterministic, and they create response probabilities through the introduction of various concepts of "error." Conceptually, they model heterogeneity in responses but not in preferences. The Fechnerian models, because they are quite specific, work most naturally with a theory that is, likewise, highly specific in its mathematical form, i.e., a model in which every component is spelled out in its full and precise functional form. They also are only well-defined if they are given a numerical hypothetical construct as input, such as the function u or the strength of preference $\mathbb{S}$ we have discussed above. Tremble models are less specific and require no numerical input; binary preference relations suffice. In that sense, tremble models are more flexible.[4]

The response models we reviewed so far have been generalized to a single broader class of "aggregation-based" specifications, according to which binary choice probabilities yield the hypothetical core deterministic preference at a suitably defined aggregate level (Regenwetter et al., 2014), such as "majority" (modal choice) or "supermajority" aggregation. Here, a hypothetical construct is only describing aggregate behavior, not necessarily every single choice made by a person. The key feature is that one or both of the following equivalences hold in tremble and Fechner models:

$$x \succ y \quad \Leftrightarrow \quad P_{xy} > \frac{1}{2} \quad \Leftrightarrow \quad u(x) > u(y). \tag{6}$$

A person is more likely to choose what he prefers than what he does not prefer. In the most general case where we consider all possible one-to-one functions u and, equivalently, all linear orders $\succ$, this representation is called the *weak utility model* (Luce and Suppes, 1965). It is equivalent to

$$\left[P_{xy} > \frac{1}{2}\right] \wedge \left[P_{yz} > \frac{1}{2}\right] \Rightarrow \left[P_{xz} > \frac{1}{2}\right] \quad \text{(for all distinct options } x, y, z\text{)}, \tag{7}$$

[3] One can also derive binary logit and probit models within a random utility framework, discussed below, by assuming that random utilities have extreme value or normal distributions, respectively.

[4] This makes them compatible with simple nonnumeric heuristics, for which Fechnerian models are ill-defined.

labeled *weak stochastic transitivity,* since the right hand side of Condition 6 forces $\succ$ in the left hand side to be transitive, and therefore Condition 7 must hold for the central term of Condition 6. Regarding the right hand side equivalence of Condition 6, it is worth noting that it only requires that one specify the function u up to a monotonic transformation. Hence, for testing, the *weak utility model* (6) is very general and inclusive. But for estimation and prediction, it is not sufficiently specific to uniquely identify the function u used in most theories.

Another class of models, whose predictions overlap with, yet also differ from, aggregation-based specifications, and which is built on different conceptual and theoretical primitives, are "random preference," "random utility," and "random function" models (Becker et al., 1963; Block and Marschak, 1960; Loomes and Sugden, 1995; Marschak, 1960; Regenwetter and Marley, 2001). These follow from the premise that the preferences and utilities, rather than the responses, are probabilistic.

In a random preference model, one considers the collection $\mathcal{R}$ of all permissible preference relations, say, for instance, $\mathcal{R}$ might denote the collection of all binary preference relations $\succ$ that are consistent with Eqn. 1 and Condition 2 using some core family of functions v, Ψ, and some core operation $\odot$, such as, say, $v(A) = A^\alpha$, $\Psi(t) = \frac{1}{1+\delta t}$, and $\times$ for $\odot$. According to such a *random preference model*, there exists a probability measure $\mathbb{P}$ on the set of all parameter values for α and δ, such that, for x giving A with time delay t and y giving B with time delay s,

$$P_{xy} = \mathbb{P}\Big(\{\alpha, \delta \mid u(x) > u(y)\} \Big) = \mathbb{P}\left(\left\{ \alpha, \delta \;\middle|\; \frac{A^\alpha}{1+\delta t} > \frac{B^\alpha}{1+\delta s} \right\} \right). \tag{8}$$

The most natural interpretation of a random preference model is that the decision maker, while fully consistent with a given core theory, is uncertain about her preferences and acts in accordance with a probability distribution over preference states that are consistent with that core theory, say, by sampling discount rates from a latent distribution. The formulation in Eqn. 8 makes it clear that this model can also be interpreted as a *random function model* (Regenwetter and Marley, 2001), since Eqn. 8 effectively makes $\mathbb{P}$ a probability measure on an appropriately defined measurable space of utility functions.

To see how much random preference models differ from tremble and Fechner models, consider, for a moment, the unusual choice between a larger, sooner and a smaller, later reward, a type of stimulus that is sometimes inserted into a study for quality control. If the respondent does not select the larger, sooner reward, this is sometimes interpreted as suggesting that he is not being attentive. Indeed, the random preference model predicts deterministic behavior in such a case because, no matter what the specific parameter values α and δ, the larger, sooner reward is preferred to the smaller, later reward: When $A > B, t < s$ in Eq. 8, then the random preference model in Eq. 8 yields $P_{xy} = 1$, regardless of the joint distribution on the values of α and δ. However, neither tremble nor Fechner models predict deterministic choice for such stimuli. Simply put, whereas a Fechner model derives probabilistic choice predictions from deterministic hypothetical constructs, a random preference model may, in certain cases, derive deterministic choice predictions from probabilistic hypothetical constructs.

A closely related *random utility model* specifies that the subjective values assigned to options x and y are uncertain. It captures this formally by defining jointly distributed random variables $\mathbf{U}_x, \mathbf{U}_y$ to denote the random utilities of options x and y. Using $\mathbb{P}$ to denote the probability measure governing the joint distribution of the random variables $\mathbf{U}_x$ (over all options x), assuming $\mathbb{P}(\mathbf{U}_x = \mathbf{U}_y) = 0, \forall x \neq y$, according to the random utility model,

$$P_{xy} = \mathbb{P}(\mathbf{U}_x > \mathbf{U}_y). \tag{9}$$

If, at every sample point of the underlying sample space, the joint realization of these random variables satisfies Conditions 1-2 with $\mathbf{U}_x$ substituted for $u(x)$, using a core family of functions $v(A) = A^\alpha$, $\Psi(t) = \frac{1}{1+\delta t}$, and $\times$ for $\odot$, then the choice probabilities in Eqns. 8 and 9 are the same. In particular, in such a random utility model, Eq. 9 gives $P_{xy} = 1$ when x is a larger sooner reward.

Just like many discounting models in the literature specify particular functions v and Φ, so do many random preference and random utility models specify properties of the probability measures $\mathbb{P}$ and/or the joint distribution of the random utilities. For example, the most commonly used random utility models assume multivariate normal distributions (probit) or extreme value distributions (logit), oftentimes for mathematical and statistical convenience. In both cases, $P_{xy} < 1$ in 'quality control' stimuli where x is a larger sooner reward. For very 'similar' stimuli, P_{xy} can, in fact, be 'close' to $\frac{1}{2}$. As we have seen earlier, these parametric random utility models are also Fechner models. However, the fully general class of random utility models makes no distributional assumptions.

2.3 Interplay between Preference, Choice, and Heterogeneity

Even just within the paradigm of models of the form $u(x) = v(A) \odot \Psi(t)$ of Eqn. 1, we face a combinatorial explosion of possible models. A fully specified model of binary choice probabilities for this paradigm states the permissible functions v and Ψ and their permissible parameter values, as well as the permissible operations $\odot$, if it is to fully detail the deterministic core hypothetical constructs. In addition, one needs to consider a suitable response mechanism, such as, e.g., upper bounds on permissible error rates ϵ_{xy}, an operation $\ominus$, a distribution function F. Or, if considering a probabilistic generalization of its core hypothetical constructs, it may need to spell out distributional assumptions about random preferences or random utilities.[5] The full range of these considerations has received little attention in intertemporal choice research because the latter has primarily focused on the algebraic core only.

For example, for monetary rewards, and $u(x) = v(A) \odot \Psi(t)$, $v(A) = A$, letting $\odot$ be the $\times$ operation, $\Psi(t) = \delta^t$, letting $\ominus$ be the $-$ operation, and F a normal CDF Φ with mean 0, we obtain a Thurstonian (aka probit) model of exponential discounting. Writing A_L, A_S for the larger and smaller rewards of L and S respectively, and t_L, t_S for the corresponding longer and shorter time delays, preference among L and S is deterministic, and responses probabilistic via

$$P_{LS} = \Phi\big(A_L\delta^{t_L} - A_S\delta^{t_S}\big). \tag{10}$$

In a random preference model, on the other hand, using the same deterministic core (but leaving out $\ominus$, which it does not use), preferences are probabilistic, and responses deterministic, via

$$P_{LS} = \mathbb{P}\big(\{\delta \mid A_L\delta^{t_L} > A_S\delta^{t_S}\}\big), \tag{11}$$

possibly with some constraints on the distribution of values of δ, say, a truncated normal distribution. Even though they are both grounded in standard exponential discounting, these two models have very different motivations: One is derived from assuming deterministic preference and probabilistic responses, the other is derived from deterministic responses based on probabilistic preferences. These models also feature drastically different mathematical properties, hence they make distinctly different predictions about behavior. In other words, not only do they make different assumptions about the source and substantive meaning of heterogeneity, they also generate different predictions about the type of heterogeneity of behavior one may observe.

Here, we are particularly interested in the types of heterogeneity different models permit. A probability mixture of models each satisfying Eqn. 10 need not, itself, satisfy Eqn. 10: Consider $0 \leq p_1, p_2, \ldots, p_k \leq 1$ with $\sum_{i=1}^k p_i = 1$ and let $\delta_1, \delta_2, \ldots, \delta_k$ be distinct parameter values. Then, there generally does not exist a parameter value δ such that

$$\Phi\big(A_L\delta^{t_L} - A_S\delta^{t_S}\big) = \sum_{i=1}^{k} p_i\Phi\big(A_L\delta_i^{t_L} - A_S\delta_i^{t_S}\big),$$

which means that tests of this model cannot let choice probabilities change/drift excessively within a person over the course of an experiment, and one cannot safely pool data across respondents who differ in their core preferences. In contrast, mixtures of models, each satisfying the distribution-free form of Eqn. 11, do, in turn, satisfy Eqn. 11: Consider $0 \leq p_1, p_2, \ldots, p_k \leq 1$ with $\sum_{i=1}^k p_i = 1$ and let $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_k$ be distinct probability measures. Then there always exists a probability measure $\mathbb{P}$ such that

$$\mathbb{P}\big(\{\delta \mid A_L\delta^{t_L} > A_S\delta^{t_S}\}\big) = \sum_{i=1}^{k} p_i\mathbb{P}_i\big(\{\delta \mid A_L\delta^{t_L} > A_S\delta^{t_S}\}\big),$$

namely, $\mathbb{P} = \sum_{i=1}^k p_i\mathbb{P}_i$. This means that these models permit high degrees of heterogeneity within and across individuals. On the other hand, distribution-free models like the one in Eqn. 11 can be mathematically intractable and most distribution-free random preference models require "order-constrained" statistical methods (Regenwetter et al., 2011, 2014).

There is, however, also much potential for model mimicry among models that are, like these, derived even from very different conceptual and mathematical primitives: While different probabilistic choice models make different predictions, it is important to note that some of their predictions usually overlap. For example, both Eqn. 10 and Eqn. 11 predict near-certain choice of L if $A_L\delta^{t_L} - A_S\delta^{t_S}$ is very large in Eqn. 10 and if Eqn. 11

[5] For prior examples of such research programs, see Stott (2006) or Blavatskyy and Pogrebna (2010). These papers considered various combinations of core theory and probabilistic specification in the domain of risky choice.

places nearly all probability mass on δ-values for which $A_L\delta^{t_L} - A_S\delta^{t_S}$ is positive. In general, however, neither Eqn. 10 implies Eqn. 11 nor vice-versa, that is, neither model is a special case of the other.

The literature on discounting models has made it quite clear that every detail about v, Ψ, and $\odot$ matters, and many papers are dedicated to discussing the details of the deterministic core structure (Doyle, 2013; Frederick et al., 2002). The literature on probabilistic response mechanisms, much of which has operated in empirical paradigms outside intertemporal choice, has likewise highlighted that every detail about probabilistic response mechanisms matters, because mis-specified response mechanisms lead to distortions of the deterministic core in statistical tests and in statistical estimation. Many papers are, in turn, dedicated to discussing the details of response mechanisms, primarily in risky choice (Birnbaum, 2011; Blavatskyy, 2011; Blavatskyy and Pogrebna, 2010; Hey, 2005; Iverson, 1990; Loomes et al., 2002; Luce, 1997; Stott, 2006; Wilcox, 2008). The intertemporal choice literature has much to gain from taking a similarly comprehensive look at sources of heterogeneity and how to model them beyond just refined deterministic cores.

Using the framework we provided above, one can select one or several specifications of hypothetical constructs, and one or several probabilistic specifications, to construct a collection of competing models of pairwise choice probabilities. One can then evaluate these competing models on suitably designed stimuli using the appropriate statistical methods. Exploring, testing, and statistically estimating every possible combination of fully specified deterministic and probabilistic components, even among a modest collection of cases like those we reviewed in the previous two subsections, poses formidable challenges: 1) Because of the many moving parts in a fully explicit theory, there can easily be thousands of combinations one may need to consider in a comprehensive analysis. 2) Models grounded in different or similar conceptual primitives need not imply the analogous similarities and differences in their probabilistic and statistical properties. 3) Different models differ strongly in their a priori flexibility to accommodate potential empirical data. 4) Parsimony in the model of hypothetical constructs can be completely disconnected from parsimony of the resulting choice model: Models with a larger number of parameters in the deterministic core need not be more flexible in their full probabilistic formulation. In fact, they can easily be more parsimonious in the space of permissible probabilistic responses. Hence, the standard approach of evaluating the parsimony of a theory by counting the number of parameters used by its deterministic functional specification is only a coarse heuristic. 5) Allowing for individual differences compounds the complexity and computational cost of reconciling preference, choice, and heterogeneity.

In light of these challenges, we proceed in a manner different from typical model selection approaches. Instead of considering specific functional forms for preferences, as is common in the literature, we abstract away to a core property shared by a large class of models for intertemporal preferences: transitivity of intertemporal preference. In other words, we follow a long tradition of axiom testing as a method to triage viable theories. Instead of considering specific functional forms of probabilistic response mechanisms, we abstract away to broad classes of probabilistic choice models. We create a collection of twenty models of pairwise choice probabilities by (1) varying whether we allow for one, some, or all transitive preferences, (2) varying whether we consider preferences, choices, or both to be probabilistic, and (3) varying the upper bounds on error probabilities where applicable. Applying these 20 models to several different stimulus sets and investigating their performance at both the individual and collective level allows us to document in detail the tradeoff between heterogeneity and parsimony.

3 Probabilistic choice models of transitive intertemporal preference

We consider twenty probabilistic choice models of transitive intertemporal preference at various levels of parsimony (see also Fig. 1). These twenty models form eight distinct model types. Four of these model types build on the theoretical premise that preferences, utilities, or strengths of preference are deterministic and that responses are probabilistic. These are the noisy-$\mathcal{P}$ (noisy patience), noisy-$\mathcal{I}$ (noisy impatience), noisy-$\mathcal{PI}$ (noisy patience-impatience), and noisy-$\mathcal{LO}$ (noisy linear order) models, each of which we consider with three different bounds on error rates. Two model types treat preferences as probabilistic and model responses as deterministic reflections of those preferences. These are the random-$\mathcal{LO}$ (random linear order) and the random-$\mathcal{LOT}$ (random linear order with tradeoffs) models. The other two model types are hybrids derived from the assumption that both preferences and responses are probabilistic. These are the noisy-$\mathcal{PI}$-mix (noisy patience-impatience mixture), and the noisy-$\mathcal{LO}$-mix (noisy linear order mixture) models, each of which we consider with three different bounds on error rates.

Insert Figure 1 here

3.1 Deterministic preferences revealed through a probabilistic response process

We first consider a simple model in which a decision maker's preference corresponds to the linear order $\succ_A$ that rank orders the choice alternatives from most to least desirable reward, no matter the time delay. A possible reason for this could be that the differences in time delays used in a given study might be perceived as negligible, compared to the relative attractiveness of the rewards. Hence, this preference ordering could derive from a more highly structured mathematical model like the general class of models (1)-(2) we reviewed earlier: For example, the functions v and Ψ of $u = v \odot \Psi$ might yield the linear order $\succ_A$ on the stimuli used in the study. For one collection of stimuli in our experimental study (our "Set 5" stimuli), this is the case, for example, when $v(A) = A, \Psi(t) = \frac{1}{1+\delta t}$ and $\odot = \times$, regardless of the discount parameter $\delta > 0$: Hyperbolic discounting makes very restrictive predictions about preferences for our Set 5. Alternatively, it could capture a simple "larger is better, no matter when" heuristic on some domain of stimuli. It is natural to suspect that the model may be limited to idiosyncratic data, i.e., it may only perform well for certain stimuli and certain respondents.

The noisy-$\mathcal{P}$ model. Suppose that possible rewards are linearly ordered. An example would be distinct cash rewards, ordered from largest to smallest amounts. The *noisy-$\mathcal{P}$ model* (noisy patience model) states that the decision maker facing L versus S chooses the larger, later reward, L, regardless of time delay, up to random error. Formally, writing $\succ_A$ for the ordering of the options from best to worst reward and setting $0 < \tau \leq \frac{1}{2}$ as *upper bound* on the permissible error rate,

$$P_{xy} \begin{cases} \geq 1 - \tau & \text{if } x \succ_A y, \\ \leq \tau & \text{if } y \succ_A x. \end{cases} \tag{12}$$

Special cases of noisy-$\mathcal{P}$: One possibility is a tremble model of $\succ_A$, according to which a decision maker has fixed preference $\succ_A$ and fixed probabilities ϵ_{xy} of making an error, with each $0 < \epsilon_{xy} < \tau$. The noisy-$\mathcal{P}$ model is more general in that only the upper bound τ on error rates is fixed, and error rates are permitted to vary, subject to the upper bound constraint. Hence, the error rates are not assumed to be statistically identifiable, nor are they assumed to be constant over time or across respondents. Alternatively, for monetary rewards, the decision maker might have a (fixed) utility function $u = v \odot \Psi$, which, when constrained to the options used in the study, happens to be monotonically increasing in the magnitude of the rewards. If L involves receiving A_L and S involves receiving only A_S, with $u > 0$, a specific Fechnerian (probit) model could state

$$P_{xy} = \Phi\left(ln\left(\frac{{A_L}^{\alpha}}{{A_S}^{\alpha}}\right)\right),$$

where Φ is a cumulative normal with mean zero. Here, the core theory models a decision maker consistent with a concave exponential utility function for money with exponent $\alpha < 1$, whose strength of preference is the ratio of subjective utilities. This model is also nested in the noisy-$\mathcal{P}$ model with $\tau = \frac{1}{2}$.

In sum, there are many possible ways to construct examples of the noisy-$\mathcal{P}$ model from either very specific or rather abstract assumptions about the subjective perception of rewards, the perception of time, the perception of the interplay between rewards and time, as well as a multitude of response mechanisms. No matter the details of such a construction, the model describes a patient decision maker with a deterministic core preference $\succ_A$ and noisy responses.

The noisy-$\mathcal{I}$ model. The *noisy-$\mathcal{I}$ model* (noisy impatience model) states that the decision maker chooses the smaller, sooner reward, S, regardless of the reward magnitude, up to random error. Formally, writing $\succ_t$ for the ordering of the options from soonest to latest, and setting $0 < \tau \leq \frac{1}{2}$ as *upper bound* on the permissible error rate,

$$P_{xy} \begin{cases} \geq 1 - \tau & \text{if } x \succ_t y, \\ \leq \tau & \text{if } y \succ_t x. \end{cases} \tag{13}$$

Note that, for any S and L pair, we have $S \succ_t L$ and $L \succ_A S$. As was the case for the noisy-$\mathcal{P}$ model, the noisy-$\mathcal{I}$ model includes a multitude of nested submodels and, hence, abstracts away from a multitude of models about subjective perceptions of rewards, time, their interaction, and response mechanisms. Despite these abstractions, this model is rather restrictive in that it only permits one single core deterministic preference relation.

The noisy-$\mathcal{PI}$ model. The noisy-$\mathcal{P}$ model and the noisy-$\mathcal{I}$ model are extreme cases where either only the linear order of the options along the dimension of the reward or the dimension of time matters. A slight generalization, allows either $\succ_A$ or $\succ_t$ to be the underlying core deterministic preference, i.e., it has a free parameter $\succ$ that may take two 'values,' namely either $\succ_A$ or $\succ_t$.

The *noisy-$\mathcal{PI}$ model* (noisy patience or impatience model) states that the decision maker is either consistently patient or consistently impatient, for a given stimulus set. More precisely, she either consistently prefers L to S, regardless of the time delays, or consistently prefers S to L, regardless of the monetary values, and chooses the preferred option up to random error. Setting $0 < \tau \leq \frac{1}{2}$ as *upper bound* on the permissible error rate,

$$\exists \succ \in \{\succ_A, \succ_t\} \quad \text{such that} \quad P_{xy} \begin{cases} \geq 1-\tau & \text{if } x \succ y, \\ \leq \tau & \text{if } y \succ x. \end{cases}$$

One attraction of this model is its potential to account for different stimulus sets in a very parsimonious fashion: A person may be patient for all stimuli in some stimulus sets and impatient for all stimuli in other stimulus sets. For example, for four of our stimulus sets, this model is a natural abstraction of hyperbolic discounting, i.e., $\Psi(t) = \frac{1}{1+\delta t}$, $v(A) = A$ and $\odot = \times$. For our experimental stimulus collections labeled "Set 1" through "Set 4," hyperbolic discounting makes very restrictive predictions: In each case, regardless of the discount parameter δ, the resulting preference is either $\succ_A$ or $\succ_t$. However, one can specify a multitude of other models that would predict either $\succ_A$ or $\succ_t$, besides hyperbolic discounting.

The noisy-$\mathcal{LO}$ model. Moving beyond patience and impatience, we also consider richer models that permit true trade-offs among reward and time. The first model of this kind permits every linear order as a core preference (or, equivalently, permits every one-to-one utility function u). Like the noisy-$\mathcal{P}$ and noisy-$\mathcal{I}$ models, it features a free parameter τ that can be interpreted as the maximal permissible error rate. With the most generous choice of error bound, $\tau = \frac{1}{2}$, this model becomes the weak utility model (6), one of the staple probabilistic models used for testing transitivity of preferences in the literature (Tversky, 1969).

The *noisy-$\mathcal{LO}$ model* (noisy linear order model) states that there exists a fixed linear order $\succ$ of the options, such that the decision maker chooses in accordance with $\succ$, up to random error. The linear order in question is unknown to the experimenter and must be inferred from the data. Formally, writing $\mathcal{LO}$ for the collection of all linear orders of the options, and setting $0 < \tau \leq \frac{1}{2}$ as *upper bound* on the permissible error rate,

$$\exists \succ \in \mathcal{LO} \quad \text{such that} \quad P_{xy} \begin{cases} \geq 1-\tau & \text{if } x \succ y, \\ \leq \tau & \text{if } y \succ x. \end{cases}$$

The noisy-$\mathcal{P}$ model and the noisy-$\mathcal{I}$ model are both nested in the noisy-$\mathcal{LO}$ model: Since $\succ_A \in \mathcal{LO}$ and $\succ_t \in \mathcal{LO}$, if a person satisfies the noisy-$\mathcal{P}$ model or the noisy-$\mathcal{I}$ model then she also satisfies the noisy-$\mathcal{LO}$ model. The noisy-$\mathcal{LO}$ model with $\tau = \frac{1}{2}$ is called "weak stochastic transitivity" (7) and the "weak utility model" (6) in the literature (Becker et al., 1963; Block and Marschak, 1960; Luce and Suppes, 1965; Marschak, 1960). Weak stochastic transitivity requires advanced order-constrained statistical methods (Iverson and Falmagne, 1985; Myung et al., 2005)[6] for a direct test. Tsai and Böckenholt (2006) tested a probabilistic intertemporal choice model on data of Roelofsma and Read (2000) and obtained choice probability estimates consistent with weak stochastic transitivity.[7] Dai (2014) tested weak stochastic transitivity directly using order-constrained Bayesian methods and found it to be well supported in an intertemporal choice task.

The noisy-$\mathcal{LO}$ model is clearly far less parsimious than the noisy-$\mathcal{P}$ model, the noisy-$\mathcal{I}$ model, or the noisy-$\mathcal{PI}$ model since it is flexible enough to permit *any linear order* as deterministic core preferences (and *any* one-to-one utility function u). On the flip-side, this may enable us to model more respondents and more types of stimuli. At the same time, however, it is important to note that this model is highly sensitive to heterogeneity: Put simply, if we randomly select decision makers who each satisfy weak stochastic transitivity, and we let them make intertemporal choices, then their overall combined (pooled) choice probabilities typically violate weak stochastic transitivity.[8] In any probabilistic choice model with deterministic core preferences, heterogeneity across individuals and/or across time is a recipe for havoc. The same problem applies to the special cases in which linear orders are derived from functional forms: If a person's parameter values within a fixed functional form for, say, a discounting model, drift over the course of an experiment, then the person's overall choice probabilities may violate the noisy-$\mathcal{LO}$ model, even though every individual choice may have originated from

[6] As Regenwetter et al. (2011) discuss in the context of risky choice, there are many published papers with inadequate tests of weak stochastic transitivity.

[7] Roelofsma and Read (2000) had interpreted their findings as evidence for intransitivity. Our R&R stimulus set uses stimuli similar to those of Roelofsma and Read (2000) to bring all 20 of our models to bear on that debate.

[8] The weak utility model's sensitivity to heterogeneous populations is historically known as the famous *Condorcet paradox* of social choice theory (Condorcet, 1785).

that model. The same applies to interindividual differences: If two decision makers satisfy, say, probit models of hyperbolic discounting (i.e., models that satisfy weak stochastic transitivity), but they use different discount rates, then their averaged choice probabilities need not satisfy a probit model of hyperbolic discounting at all, and typically do not even satisfy weak stochastic transitivity.[9]

3.2 Probabilistic preferences revealed through a deterministic response process

Random preference and certain distribution-free random utility models start from fundamentally different premises than the four models we have just discussed. Here, the decision maker is uncertain about which option is preferable, yet, no matter which sample point of the underlying sample space is realized, the core theory is fully satisfied. Conditional on the momentary preference, the response is error-free.

The random-$\mathcal{LO}$ model. Binary choice probabilities satisfy the *random-$\mathcal{LO}$ model* (random linear order model) if there exists a probability distribution over linear orders such that the binary choice probability of choosing L over S is the total probability of those linear orders in which L is preferred to S. Formally, let $\mathcal{LO}$ denote the collection of all linear orders on a given set of choice options. Binary choice probabilities satisfy the *random-$\mathcal{LO}$ model* if there exists a probability distribution $\mathbb{P}$ on $\mathcal{LO}$, that is, $0 \leq \mathbb{P}(\succ) \leq 1, \forall \succ \in \mathcal{LO}$ and $\sum_{\succ \in \mathcal{LO}} \mathbb{P}(\succ) = 1$, such that

$$P_{xy} = \sum_{\substack{\succ \in \mathcal{LO} \\ x \succ y}} \mathbb{P}(\succ) \quad \text{(for all distinct options } x, y).$$

This model is mathematically equivalent to the distribution-free random utility model (9) in that binary choice probabilities satisfy one model if and only if they satisfy the other (Block and Marschak, 1960).

The random-$\mathcal{LOT}$ model. We consider one more random preference model, namely the case in which all linear orders, except $\succ_A$ and $\succ_t$ are permissible preferences states. This model rules out the extreme cases of completely patient or completely impatient preference states. Let $\mathcal{LOT}$ denote the collection of all linear orders on a given set of choice options, except $\succ_A$ and $\succ_t$, i.e., $\mathcal{LOT} = \mathcal{LO} \backslash \{\succ_A, \succ_t\}$. Binary choice probabilities satisfy the *random-$\mathcal{LOT}$ model* (random linear order with tradeoffs model) if there exists a probability distribution $\mathbb{P}$ on $\mathcal{LOT}$, such that

$$P_{xy} = \sum_{\substack{\succ \in \mathcal{LOT} \\ x \succ y}} \mathbb{P}(\succ) \quad \text{(for all distinct options } x, y). \tag{14}$$

This model can also be restated in random utility terms. Binary choice probabilities satisfy Eqn. 14 if and only if there exist jointly distributed random variables, with $\mathbf{U}_x$ denoting the random utility of option x and $\mathbb{P}$ denoting the probability measure governing the joint distribution, with $\mathbb{P}(\mathbf{U}_x = \mathbf{U}_y) = 0, \forall x \neq y$, such that $\mathbb{P}\left(\bigcap_{\substack{r,s \\ r \succ_A s}} \mathbf{U}_r > \mathbf{U}_s\right) = 0$ and $\mathbb{P}\left(\bigcap_{\substack{v,w \\ v \succ_t w}} \mathbf{U}_v > \mathbf{U}_w\right) = 0$.

3.3 Probabilistic preferences compounded with probabilistic responses

We now consider a hybrid between the noisy-$\mathcal{P}$ model and the noisy-$\mathcal{I}$ model, and a hybrid of the random-$\mathcal{LO}$ model and the noisy-$\mathcal{LO}$ model. They follow from the general theoretical premise that preferences and responses are both probabilistic. Within an individual, this premise can capture the idea that the individual is both uncertain about his preference and responds in a noisy fashion. At the group level, these models describe a heterogeneous population of up to three types of decision makers: those with deterministic preferences who respond in a noisy fashion, those with uncertain preferences who respond in a deterministic fashion, and those with uncertain preferences who also respond noisily. We limit ourselves to the two extreme cases where either only the two preferences $\succ_A$ and $\succ_t$ are permissible, or where all linear orders are permissible.

The noisy-$\mathcal{PI}$-mix model. Let $0 < \tau \leq \frac{1}{2}$ be an *upper bound* on the permissible error rate. Let P^A_{xy} denote the binary choice probabilities according to the noisy-$\mathcal{P}$ model (12) and let P^t_{xy}denote the binary choice probabilities according to the noisy-$\mathcal{I}$ model (13). According to the *noisy-$\mathcal{PI}$-mix model* (noisy patience-impatience mixture model), there exists a mixture probability p such that, in any given pairwise choice between L and S the person

[9] These observations follow trivially from the convexity or nonconvexity of various probability spaces.

chooses according to the noisy-$\mathcal{I}$ model with probability p and according to the noisy-$\mathcal{P}$ model otherwise.[10]

$$\exists p \in [0,1] \quad \text{such that} \quad P_{xy} = pP^t_{xy} + (1-p)P^A_{xy},$$
$$\text{where } P^t_{xy} \begin{cases} \geq 1-\tau & \text{if } x \succ_t y, \\ \leq \tau & \text{if } y \succ_t x, \end{cases} \quad \text{and} \quad P^A_{xy} \begin{cases} \geq 1-\tau & \text{if } x \succ_A y, \\ \leq \tau & \text{if } y \succ_A x, \end{cases} \quad \text{(for all distinct options } x, y).$$

This model could, for example, model a population consisting of patient and impatient individuals only, with each decision maker also potentially making errors in his choices. Within person, it can model an individual who, for example, waivers between being patient and impatient, compounded with errors in her choices. This model is particularly interesting in that it does not connect to, say, discounting models, as easily as others. In order to satisfy this model, a population would have to consist of individuals whose discount rates are consistent with only the two preference rankings $\succ_A$ and $\succ_t$ on a given set of stimuli. As a discounting model of an individual, this would only allow discount rates according to which the individual either has preference $\succ_A$ or $\succ_t$. In our stimulus sets Set 1 - Set 5 (but not R&R), this is, indeed the case for hyperbolic discounting: As we have seen earlier, hyperbolic discounting predicts $\succ_A$ or $\succ_t$ regardless of discount rate in those five stimulus sets. Other discounting models predict a larger variety of preferences.

The noisy-$\mathcal{LO}$-mix model. Our most complex (i.e., least statistically parsimonious) model permits a probability distribution over all possible linear order core preferences, compounded with noisy responses. Let $0 < \tau \leq \frac{1}{2}$ be an *upper bound* on the permissible error rate, and $\forall \succ \in \mathcal{LO}$, let $p_\succ$ denote the probability of making choices according to a noisy process with $\succ$ as core preference. Then the *noisy-$\mathcal{LO}$-mix model* (noisy linear order mixture model) states that

$$P_{xy} = \sum_{\succ \in \mathcal{LO}} p_\succ P^\succ_{xy} \quad \text{with} \quad P^\succ_{xy} \begin{cases} \geq 1-\tau & \text{if } x \succ y, \\ \leq \tau & \text{if } y \succ x, \end{cases} \quad \text{(for all distinct options } x, y).$$

The noisy-$\mathcal{PI}$-mix model is a nested submodel of the noisy-$\mathcal{LO}$-mix model, in which $p_{\succ_A} = p = 1 - p_{\succ_t}$ and $P^{\succ_A} = P^A$, as well as $P^{\succ_t} = P^t$.

3.4 Summary of models

Figure 1 visualizes some of the similarities and differences between these models. Suppose that L is larger and later than M, which is, in turn, larger and later than S. The coordinates of the 3D figure show binary choice probabilities P_{MS} on the vertical axis marked (M, S), P_{LM} on the axis marked (L, M) from the origin to the right, and P_{LS} on the axis marked (L, S) from the origin to the left. The deterministic core preferences correspond to corners (binary choice probabilities equaling 0 or 1) of the 3D cube. Despite being based on similar core premises about the hypothetical constructs of preferences or utilities, the models differ dramatically in their behavioral predictions. At the same time, probabilistic choice models that are built on different underlying premises overlap in complex ways. While the figure shows correctly which models are nested (such as noisy-$\mathcal{PI}$ in noisy-$\mathcal{PI}$-mix), it is important not to over-interpret the 3D visualization with respect to the parsimony of these models. Some of the models that appear to be relatively large in Figure 1 (such as random-$\mathcal{LO}$) rapidly become very restrictive in higher dimensions (i.e., they become more parsimonious when there are more than three choice probabilities). Likewise, some models that are very restrictive on just three choice probabilities may be less so in higher dimensions (e.g. random-$\mathcal{LOT}$ is only slightly more restrictive than random-$\mathcal{LO}$ in higher dimensions).

Insert Table 1 here

Table 1 summarizes our models from a different perspective. The first column lists the model names, whereas the second column shows the set of core preference states permitted by the core theory in each model. In addition to the eight models above, we also consider a *saturated* model that places no constraints whatsoever on binary choice probabilities. Its core theory is unconstrained in that it allows all (asymmetric) binary preference relations as preference states. We denote the set of all such binary preferences by $\mathcal{B}$. Columns 4 and 5 of Table 1 summarize whether preferences and responses are each deterministic or probabilistic. The last column gives each

[10] Note that our formulation of this model does not permit p to vary with xy. However, because it forms a convex set, the model does allow some variation of p over time, including some degree of variation over repeated observations. Likewise, viewed as a model of a population, it allows for inter-individual heterogeneity in the value of p.

model a label that we use in our data analyses below. Models derived from probabilistic core preferences are shaded with a gray background. Models with deterministic response processes are marked in bold.

4 Model specification for Bayesian statistical analysis

The premise of this paper is threefold: 1) There are many moving parts to a fully specified model of intertemporal binary choice behavior, with much prior work discussing only unobservable hypothetical constructs in detail. 2) Different transitive models of observable intertemporal choice behavior vary in their parsimony. 3) We expect a tradeoff between the parsimony of a model and the variety of individuals and stimuli for which it can account, with the most parsimonious models likely working only for specific individuals and specific stimuli, and a universal model for all individuals and stimuli likely requiring extreme flexibility. In line with these conceptual expectations, we analyze our data from multiple perspectives. In contrast with most of the literature, our analyses are custom-designed to account formally for various levels and types of heterogeneity and parsimony.

We report all our analyses in Bayesian terms here and provide frequentist (hypothesis testing) analyses in the Supplementary Materials[11]. We use Bayesian p-values (Gelman et al., 1996) to assess model viability, Bayes factors (Kass and Raftery, 1995) to compare models at the level of each individual respondent, and group Bayes factors (Stephan et al., 2007) to aggregate Bayes factors across respondents. The magnitude of the Bayes factor between two models is the degree of evidence in favor of one model over the other. Our application of these methods to behavioral data follows similar recent analyses in the context of risky choice (Cavagnaro and Davis-Stober, 2014; Davis-Stober et al., 2015; Guo and Regenwetter, 2014). In those studies, as in ours, models were defined through systems of linear inequality constraints on binary choice probabilities. Because Bayesian model selection requires that, in addition to constraints on choice probabilities such as those visualized in Fig. 1, the models be cast via a likelihood function and a prior, we reformulate each set of inequality constraints using a prior distribution with support over only those probability vectors that are consistent with the model in question (see also Myung et al., 2005).

Formally, let $\mathcal{C}$ denote a collection of d distinct unordered pairs of choice options. For each pair $\{x, y\} \in \mathcal{C}$, let P_{xy} denote the binary choice probability of x being chosen from $\{x, y\}$, and let $\vec{P} = \{P_{xy}\}_{\{x,y\}\in\mathcal{C}}$ denote a 'binary choice probability vector' (because each $P_{xy} = 1 - P_{yx}$, we only use/count one of these two probabilities for each pair $\{x, y\}$). Then, for each model q defined above, let $\Lambda_q \subseteq [0, 1]^d$ denote the subset of binary choice probability vectors $\vec{P}$ satisfying the inequality constraints that characterize model q, and let v_q denote the Lebesgue measure (i.e., volume) of Λ_q. We construct the Bayesian model M_q with a uniform prior over the model, that is, with the order-constrained prior distribution

$$\pi(\vec{P}|M_q) = \begin{cases} \frac{1}{v_q} & \text{if } \vec{P} \in \Lambda_q, \\ 0 & \text{otherwise,} \end{cases} \qquad (\text{for all } \vec{P} \in [0, 1]^d).$$

Fully specified Bayesian models follow naturally by combining each order-constrained prior with a likelihood function, defined as follows. Let N_{xy} denote the number of times that the pair of delayed rewards $\{x, y\}$ is presented to the decision maker, let n_{xy} denote the number of times that x was chosen from $\{x, y\}$, and let $\vec{n} = \{n_{xy}\}_{\{x,y\}\in\mathcal{C}}$. Assuming that repeated choices from each option pair are identically distributed and that all choices are mutually independent[12], the likelihood function f for a set of responses $\vec{n}$ takes the following, product-of-binomials form:

$$f(\vec{P}|\vec{n}) = \prod_{x,y\in\mathcal{C}} \binom{N_{xy}}{n_{xy}} P_{xy}^{n_{xy}} (1 - P_{xy})^{N_{xy}-n_{xy}}. \tag{15}$$

In addition to the models we have already described, we also define a "saturated" model to serve as a common baseline against which to compare each substantive model. This model puts no constraints on binary choice probabilities, so it is defined by the prior $\pi(\vec{P}|$ saturated model$) = 1$, $\vec{P} \in [0, 1]^d$; that is, a uniform prior over the entire space of all choice probability vectors. This model is vacuous in the sense that it is guaranteed to fit any set of data perfectly. In model selection analyses that penalize for complexity, this model will receive the largest

[11] Wherever both statistical approaches are applicable, our Bayesian and frequentist analyses are well aligned in the scientific conclusions that they support. The Bayesian approach is advantageous here: It naturally handles a situation like ours, in which some but not all models are nested within each other, and some models differ strongly in their parsimony despite having the same number of free parameters (here each model is characterized by 10 Binomials).

[12] In a Bayesian framework, the same likelihood function can be derived from different theoretical primitives about the data generating process and the interpretation of P_{xy}. In particular, one may assume that repeated choices on the same option pair are infinitely exchangeable and that choices on different choices pairs are independent. See Bernardo (1996) for discussion.

penalty because it is maximally complex. The saturated model provides a common benchmark for measuring the degree of evidence supporting or contradicting each substantive model. It also lets us define what it means for a substantive model to fail: If a model's Bayes factor against the saturated model is less than 1.0, then we are better off using the saturated model (i.e., no model) than the substantive model. If the Bayes factors of all our substantive models were less than 1.0, this would suggest that the data violated a fundamental assumption shared by these models, such as, e.g., transitivity.

5 Experiment

We ran two studies aimed at evaluating the eight types of probabilistic choice models of transitive intertemporal preference. Decision makers made pairwise choices between larger, later and smaller, sooner options. The experiments were run in two locations: Urbana-Champaign in Illinois (USA) and Berlin (Germany). In each location, we used six different stimulus sets to cover a range of different stimuli. One experiment collected enough repeated choices for the same stimuli from each person (mixed with a large number of distractors) to permit individual subject analyses. The other experiment drastically simplified the task by asking each respondent to make each pairwise choice only once. Hence, the second experiment does not provide enough data from each respondent for individual-level analyses.

5.1 Respondents

Respondent recruitment and testing took place at both the University of Illinois at Urbana-Champaign (UIUC), and the Max Planck Institute for Human Development (MPI). UIUC respondents were university students and local residents. MPI respondents attended a German university and chose to participate through their institute's experimental respondent pool. All respondents received monetary rewards based on choices they made during the experiment and they only learned their reward amount after completion of the experiment. In accord with payment standards at the University of Illinois, UIUC respondents also received an additional base payment ($12 for Experiment 1 and $8 for Experiment 2).

Before experimental testing, we selected a subset of trials from which all rewards would be paid. These preselected trials all had relatively high reward amounts, thus ensuring sufficient remuneration. Each respondent's particular reward was determined by randomly selecting one of these pre-selected trials. Respondents were not informed about the mechanism by which we selected stimuli that were used for payment and whether this selection was made before or after data collection. Respondents were explicitly instructed at the beginning of the experiment to make choices based on their true preferences because they would receive one of their chosen time-delayed rewards as a real payment. We paid UIUC respondents with the exact delay specified (even if the date fell on a weekend or holiday) by implementing a payment system via an agreement between the university and Amazon.com. After the experiment was over, respondents provided an email address to which an electronic Amazon gift code (matching the U.S. Dollar value of their chosen reward) was sent on the specified calendar day in the future (matching the delay of their reward). The MPI offered respondents two options at the end of the study. If the real reward was an option that included a positive time delay, respondents could opt to receive 85% of the amount in cash immediately instead of waiting for the delayed full reward. Respondents were not told that they could substitute this immediate payment until they had completed all choices. If they opted for the full delayed reward, they received it after the specified delay through a bank transfer in euros.[13]

For Experiment 1 (at UIUC), we tested 31 respondents (14 males, 17 females) from June-October 2012 with a mean±SD age of 20.8±2.4 years (range 18-28). At MPI, we tested 30 respondents (16 males, 14 females) from June-July 2012 with a mean±SD age of 25.6±3.7 years (range 20-34). For Experiment 2 (at UIUC), we tested 34 respondents from September-November 2013. Age and gender of these respondents was not recorded. At MPI, we tested 30 respondents (10 males, 20 females) from November-December 2013 with a mean±SD age of 25.3±2.6 years (range 20-30).

5.2 Experimental procedure

The UIUC Institutional Review Board and the Ethics Committee of the MPI reviewed and approved both experiments.[14]

[13] The Supplementary Materials provide the instructions to respondents and the stimuli used for real payment.

[14] University of Illinois at Urbana-Champaign, IRB approval #11427.

5.2.1 Procedure

Respondents completed the experiments on computers. UIUC respondents saw English text and U.S. dollars for currency, whereas MPI respondents saw German text and euros for currency but identical numbers as did the U.S. respondents (not currency-converted values). Respondents could first provide their age, gender, and occupation.[15] They then read one set of instructions, completed 10 practice trials, and then read a final set of instructions before beginning the actual trials. This final instruction set informed each respondent that their reward at the end of the experiment would be determined by one of the choices made during the study. For each trial, the respondents used a computer mouse to select one of two options presented on the screen, each characterized by a specified reward amount and a time delay. At the end of the experiment, respondents were then shown the reward that they were going to receive.[16]

Experiment 1 consisted of two sessions with 1,006 trials each (including 6 warm-up trials). At UIUC, respondents completed the two sessions of Experiment 1 on two different days. MPI respondents completed the two sessions for Experiment 1 on a single day, separated by a 5-15 minute break. Each session of Experiment 1 took respondents 30-90 minutes to complete. While Experiment 1 was designed to elicit enough information from each person to permit within-respondent data analyses, Experiment 2 was aimed at collecting the same kind of data with a much smaller number of trials, for a joint analysis of all respondents combined. It had a single session with 106 questions (6 of them warm-up trials) and took respondents 10-30 minutes to complete. Respondents in Experiment 2 saw the same questions as respondents in Experiment 1, except that none of the items were repeated.

5.2.2 Stimuli

We created six option sets. Sets 1 - 5 each consisted of five intertemporal options (top of Table 2). The sets varied in the magnitude and spread of monetary amounts (stated in $ and €) and in the magnitude and spread of time delays (stated in days). For each set of five options, we created all 10 possible pairwise combinations of options to create 10 different *option pairs* per option set. Across all five sets of stimuli, this resulted in a total of 50 option pairs. We also used an additional collection of nine option pairs. We adapted one triple from stimuli in Roelofsma and Read's (2000) study of intransitivity in intertemporal choice, and two additional such triples were similar but varied and expanded the range of reward amounts. This sixth stimulus set of nine option pairs is labeled R&R (bottom of Table 2).

Insert Table 2 here

For Experiment 1, to permit within-respondent statistical analysis, respondents saw each of the 59 option pairs 20 times[17], yielding 1180 experimental trials. These 1180 trials were mixed with another 832 pairs of stimuli, some of which were designed to test other hypotheses while others served as distractors. The 2,012 pairs of stimuli were divided into *blocks*, each consisting of a series of five consecutive option pairs. Within each block, we randomized the order of presentation across respondents. The order of the blocks was constant across respondents. Each block contained two or three experimental pairs, but never from the same stimulus set. We placed option pairs from the same set in alternating blocks, so respondents saw 5-13 other pairs between experimental pairs from the same stimulus set. Respondents were shown 95-103 option pairs before experiencing a repetition of the same pair.

It is natural to question whether making in excess of 1,000 decisions per session could bias a respondent's behavior and yield unrealistic data. We tested this concern empirical by running a second experiment with the same stimuli, but with a small number of individual trials per person. Hence, for Experiment 2, where we did not aim to carry out individual respondent statistical analyses, respondents saw each of the 59 option pairs exactly once. They were also given another 47 distractor pairs. The method of sequencing the presentations of these option pairs matched that of Experiment 1.

[15] This step was accidentally omitted by the person administering Experiment 2 at UIUC.

[16] The experimental software used was a custom-made program called *Disc'n'Risk*, developed by Uwe Czienskowski at MPI. The Supplementary Materials give further experimental details.

[17] We repeated each option pair 20 times in order to accommodate a frequentist analysis. If we only ran the Bayesian analysis, we could cut this by a factor of 3. For example, Davis-Stober et al. (2015) used 8 repetitions per option pair in a ternary choice experiment. Some parametric models, such as logit and probit models work without repetitions of the same stimuli and, instead, use many different stimuli for statistical convergence.

6 Results of Experiment 1

We tested all eight model types of Section 3, as illustrated three-dimensionally in Figure 1. For noisy-$\mathcal{P}$, noisy-$\mathcal{I}$, noisy-$\mathcal{PI}$, noisy-$\mathcal{PI}$-mix, noisy-$\mathcal{LO}$, noisy-$\mathcal{LO}$-mix, we furthermore used three different bounds τ on error rates: $\tau = 0.5$ (modal choice, which contains Fechnerian models, such as logit and probit specifications, as special cases), $\tau = 0.25$ (whose maximum error rate is considered adequate by some scholars, e.g., Harless and Camerer 1994), and $\tau = 0.1$ (according to which errors are not a major component of the response process). All in all, therefore, we tested 20 different transitive probabilistic models of intertemporal choice. All of our analyses require order-constrained statistical inference, implemented in the public domain software QTest, programmed for multiple computing platforms[18] (Regenwetter et al., 2014).

6.1 Are transitive models viable?

We first assess the overall viability of each model for each respondent and stimulus set by computing the Bayesian p-values (Gelman et al., 1996). The Bayesian p-value is a posterior predictive check of the descriptive adequacy of each model. It is computationally inexpensive and relatively easy to interpret. Essentially, the Bayesian p-value is computed by comparing the observed data to the posterior predictive distribution of the model. If the observed data are consistent with the posterior predictive distribution, then the Bayesian p-value is high; otherwise, it is low (see Myung et al., 2005, for details on computation). A standard approach is to declare an *adequate fit* of a model to the data whenever the Bayesian p-value exceeds 0.05. The Bayesian p-value does not indicate the probability that a model is correct. Bayesian p-values cannot be compared across models. We use Bayesian p-values only to determine the proportion of respondents for whom each model provides at least an adequate fit, and we leave model selection for later.

We computed the Bayesian p-value of each model separately for each respondent and stimulus set. Figure 2 shows, for each model and stimulus set, the proportion of respondents for whom that model provided an adequate fit (frequentist fits are available in Figure S1 in the Supplementary Materials). Overall, there seem to be several transitive models that provide adequate fits for most respondents and most stimulus sets. The most complex model, in which all linear orders are permissible preference states and in which responses can be maximally noisy, the noisy-$\mathcal{LO}$-mix model with $\tau = \frac{1}{2}$, provides an adequate fit for nearly every respondent in every stimulus set. On the one hand, this means that transitive models can account almost universally for our data across respondents and stimulus sets. On the other hand, the three instances of the noisy-$\mathcal{LO}$-mix are among the most statistically complex of the models we have tested, and the Bayesian p-value does not penalize models for complexity.

In contrast, the noisy-$\mathcal{I}$ models at all noise bounds were inadequate for all but a few respondents in each stimulus set, casting doubt on this model's viability as an explanation of the data at any level of the error bounds. However, since this model is especially parsimonious relative to the others, especially at the 0.1 noise bound, it is possible that a noisy-$\mathcal{I}$ model could provide the best explanation for those respondents and stimulus sets in which its Bayesian p-value exceeded 0.05.

The random-$\mathcal{LO}$ model fits a large proportion of respondents. When the $\succ_A$ and $\succ_t$ options are removed in the random-$\mathcal{LOT}$ model, however, the fit drops dramatically. The large decrease in fit caused by the removal of these preference states suggests that linear orders based exclusively on either amount or time played a key role in the good performance of the random-$\mathcal{LO}$ model.

The noisy-$\mathcal{P}$ models seem to show the greatest interaction across stimulus sets, especially at the 0.25 and 0.1 noise bounds, as they are adequate for most respondents in Stimulus Sets 3, 4, and 5, but fewer than half of the respondents in Sets 1, 2, and R&R. Similar patterns of interaction emerge for the noisy $\mathcal{LO}$ models, noisy-$\mathcal{PI}$ models, and noisy-$\mathcal{PI}$-mix models, especially those with lower error bounds τ. These results raise the question whether respondents' behavior may be best described by different models in different stimulus sets, with an overall model across stimulus sets requiring some flexibility. To answer this question more conclusively, we proceed to the model selection analysis.

Insert Figure 2 here

[18] The original (frequentist only) release of QTest is available at www.regenwetterlab.org. A new multicore compatible version with Bayesian capabilities is available from the authors while it is being prepared for public release.

6.2 Model selection results: Individual level analyses

Our next goal is to identify the best model at the individual level, before we proceed to the group level. Our criterion for model selection is the Bayes factor (Kass and Raftery, 1995), defined as the ratio of the marginal likelihoods of two models, derived from Bayesian updating. The Bayes factor accounts for both goodness-of-fit and complexity/parsimony. It selects among models based on generalizability (Pitt and Myung, 2002), in that the model with the highest Bayes factor is the one deemed to most accurately predict future data samples from the same process that generated the currently observed sample (see, e.g., Liu and Aitkin, 2008).

To identify the best model at the individual level, we computed the Bayes factor of each model, relative to the saturated model, separately for each respondent and stimulus set.[19] With 20 models, 61 respondents, and 6 stimulus sets in our study, this analysis yielded a total of 7,320 respondent-level Bayes factors. Our Bayes factors varied across many orders of magnitude (the Bayes factors for each model, respondent, and stimulus set are available in a spreadsheet that is part of the Supplementary Materials). Many Bayes factors were quite large and, hence, provided strong evidence in favor of the model in question. However, likewise, in many cases, the evidence against a given model was quite strong: Of the nearly 3,000 Bayes factors that were smaller than 1.0, nearly half (1,450) had log_{10} values between -10 and -200. Of these, 984 were for the noisy-$\mathcal{I}$, 223 were for the noisy-$\mathcal{P}$, 131 were for the noisy-$\mathcal{PI}$, 58 were for the noisy-$\mathcal{PI}$-mix and 54 were for the noisy-$\mathcal{LO}$. Table 3 summarizes the results by reporting key features of the best model for each respondent and stimulus set. The features are identified using the labels introduced in Table 1. For example, the best model for Respondent 1 in Set 1 in the UIUC sample is noisy-$\mathcal{PI}$-mix, which assumes probabilistic preferences and choices. So, the corresponding cell is shaded to indicate probabilistic preferences and it shows the core theory $\{\succ_A, \succ_t\}$ in plain text (rather than bold) to indicate probabilistic choices. For simplicity, the table uses the same label for all models with the same core theory, preferences, and response process, regardless of error bound (e.g., noisy-$\mathcal{PI}$-mix with $\tau = 0.5$ and noisy-$\mathcal{PI}$-mix with $\tau = 0.1$).

Insert Table 3 here

Perhaps the most prominent aspect of Table 3 is the apparent heterogeneity across respondents and stimulus sets. No single core theory, type of preference, or type of response process was robust across all respondents and stimulus sets. In fact, not only was there heterogeneity in terms of the best model, there was also heterogeneity in terms of which models were adequate. That is to say, no model had a Bayes factor greater than 1.0 for every respondent and stimulus set, meaning that every model failed on at least one respondent and stimulus set, relative to the saturated model (see the spreadsheet in the Supplementary Materials for the Bayes factor of each model, respondent, and stimulus set). This does *not* mean all of the models failed overall, as there were only very few cases (8 out of 366 respondent-by-stimulus combinations, indicated by the black shaded boxes in Table 3) in which none of the 20 models had a Bayes factor greater than 1.0. Nevertheless, the 8 cases in which the saturated model was favored represent instances in which transitivity (a core assumption shared by all 20 models under consideration) may have been violated. In the current modeling framework, a violation of transitivity means that the core theory of the best model includes one or more intransitive preferences. Interestingly, four of the apparent violations involved just two respondents: UIUC Respondent 14 and MPI Respondent 22; and six of them involved just one stimulus set: Set 2. This clustering of apparent violations within certain experimental conditions and respondents is consistent with the findings of Cavagnaro and Davis-Stober (2014) and suggests that the violations may represent robust individual differences.

Although no core theory was best across the board, $\succ_A$ most frequently performed best (264 of 366 entries in Table 3), indicating that most respondents seem to prefer the option with the highest amount, regardless of the time delay. This was especially the case in Stimulus Sets 4 and 5, in which all but eight respondents were best described by a model assuming core theory $\succ_A$. In contrast, fewer than two-thirds of respondents were best described by $\succ_A$ in Stimulus Sets 1, 2, and R&R. Despite these variations across stimulus sets, we found that about half (31 out of 61) respondents were best described by the same core in all six stimulus sets (these are marked in Table 3 with respondent numbers enclosed in hyphens, e.g., -2-). This consistency suggests that the

[19] In general, Bayes factors of inequality constrained models cannot be obtained analytically. However, in this particular case, we were able to obtain analytical solutions for the Bayes factors of noisy-$\mathcal{P}$, noisy-$\mathcal{I}$, noisy-$\mathcal{PI}$, and noisy-$\mathcal{LO}$, relative to the saturated model. This is because the inequality constraints are orthogonal within each of these models, and the priors on each dimension are independent and conjugate to the likelihood function. We obtained respondent-level Bayes factors for the remaining models, in which the order constraints are not orthogonal, using Monte Carlo integration. To compute pooled Bayes factors, we used a specialized procedure developed by Klugkist and Hoijtink (2007). In short, this algorithm yields the Bayes factor for an order-constrained model versus the saturated model by drawing a large sample from the posterior distribution of the saturated model and computing the proportion of the sample that satisfies the order constraints of the nested model (see Cavagnaro and Davis-Stober, 2014, for additional details).

best core theory may be somewhat robust across stimulus sets, within some respondents.

Like any model selection analysis on experimental data, our analysis is specific to the models, participants, and stimuli considered. The fact that $\succ_A$ accounts well for some stimulus sets but not others suggests that it is worthwhile considering core theories that agree with $\succ_A$ on some stimulus sets but not others. In our Roadmap section, we discuss how to evaluate a variety of core theories using the same general approach, and with appropriate stimuli.

6.3 Model selection results: Group level analyses

To select among models at the group level, we use two measures: the group Bayes factor (GBF, Stephan et al., 2007) and the pooled Bayes factor (PBF). Both select among models at the group level, but they differ in the mechanism by which respondent-level results are aggregated: the PBF aggregates *data* across respondents, whereas the GBF aggregates *likelihoods* across respondents. The PBF is the ratio of the marginal likelihoods of two models given the pooled data from all respondents, whereas the GBF is the product of respondent-level Bayes factors. Thus, the model with the highest PBF is the one that best accounts for the pooled data, while the model with the highest GBF is the one that *jointly* best accounts for each respondent's data.[20]

Table 4 ranks each model based on the GBF and PBF, respectively, in each stimulus set (the log_{10} transformed GBF and PBF values are reported in Table S2). For pooled data, it only makes sense to evaluate models which, if there is more than one core deterministic preference, can inherently accommodate heterogeneity of preferences. Formally, these are models whose parameter spaces form convex sets, i.e., we must omit the noisy-$\mathcal{PI}$ model and the noisy-$\mathcal{LO}$ model (such as weak stochastic transitivity).

The noisy-$\mathcal{PI}$-mix model was by far the most successful, according to both the GBF and PBF, in almost all stimulus sets. The exceptions were Set 2, in which noisy-$\mathcal{LO}$ was best according to GBF, and R&R, in which noisy-$\mathcal{LO}$ and noisy-$\mathcal{P}$ were best according to the GBF and PBF, respectively. What is most notable about this result, besides the near-unanimity across stimulus sets, is that noisy-$\mathcal{PI}$-mix assumes probabilistic preferences, whereas a vast majority of respondents were best described as having deterministic preferences. These results are not contradictory, as they may seem at first, because probabilistic preferences at the group level need not imply that every decision maker in the group has uncertain preferences. Rather, probabilistic preferences at the group level implies that the sample comprises a heterogeneous mix of up to three types of decision makers: those with deterministic preferences who respond in a noisy fashion, those who have uncertain preferences and respond in a deterministic fashion, and those who have uncertain preferences and respond in a noisy fashion. The group-level analyses cannot identify the nature of the heterogeneity more precisely because they do not distinguish between variability within respondents (such as, preference uncertainty) and variability between respondents (such as, individual differences in core preferences).

Insert Table 4 here

Despite the limitations of the group-level analyses, they are essential for obtaining results that generalize beyond each particular decision maker. The current GBF results suggest that the model that will generalize best to data from a randomly selected respondent is noisy-$\mathcal{PI}$-mix. Although this model implies probabilistic preferences $\succ_A$ and $\succ_t$, we can see from the respondent-level results, in Table 3, that it is unlikely for a randomly selected respondent to be best described by such a model (most are best described by models with deterministic preference $\succ_A$). However, since there are individual differences, the randomly selected respondent may be best described as having deterministic preference $\succ_A$, or deterministic preference $\succ_t$, both of which are part of $\succ_A \vee \succ_t$. Thus, noisy-$\mathcal{PI}$-mix is selected by the GBF because it is deemed to provide the most parsimonious account that is consistent with the behavior of most respondents.

It also stands out that noisy-$\mathcal{PI}$-mix does well in only four of the six stimulus sets, whereas noisy-$\mathcal{LO}$ does well in Set 2 and R&R. In fact, the only models that beat the unconstrained model across all six stimulus sets are noisy-$\mathcal{LO}$ with error rates of 0.25 and 0.5. This suggests that generalizing across multiple stimulus sets requires more preference patterns than just $\succ_A$ and $\succ_t$. This result highlights the importance of the choice of stimulus sets when testing models of intertemporal choice. If one is only concerned with modeling choices on a narrow set of stimuli, such as those in Sets 3-5, then a small set of preference patterns may suffice. However, generalizing to a broader set of stimuli may require additional preference patterns, perhaps even intransitive patterns. Identifying the minimal set of preference patterns that generalizes to any arbitrary stimulus sets is

[20] This interpretation of the GBF rests on two assumptions: that every respondent has the same model (i.e., the same set of restrictions on choice probabilities, but not necessarily the same choice probabilities) and that the model evidences are independent. The latter assumption is tenable for GBFs as long as respondents are sampled independently from the population.

beyond the scope of this paper. Later, in the Roadmap section, we provide additional guidance on investigating this issue.

The pooled Bayes factor results suggest a slightly different interpretation than the group Bayes factor results. Since the PBF is based on pooled data, the model selected by the PBF is the one that is deemed to generalize best to future pooled data. That is, it may not be representative of any particular respondent, but it parsimoniously captures the aggregate choice proportions. This distinction between the PBF and the GBF helps to explain why the noisy-$\mathcal{P}$ model fares well according to the PBF but not the GBF. The noisy-$\mathcal{P}$ model fares well according to the PBF because, in the pooled data, any influence from the minority of respondents whose choices are not consistent with noisy-$\mathcal{P}$ (i.e., those in Table 3 whose best core theory was not $\succ_A$) is washed out by the vast majority of respondents whose choices are best described by noisy-$\mathcal{P}$. On the other hand, the GBF is not based on pooled data, but rather aims to simultaneously describe each respondent's choice proportions. Thus, the noisy-$\mathcal{P}$ model does not fare well according to the GBF, because the noisy-$\mathcal{P}$ model provides such an extremely poor account of the choice data from those respondents who were best described by other models (e.g., those in Table 3 whose best core theory was $\succ_t$).

7 Results of Experiment 2

Experiment 2 aimed to diagnose systematic changes in respondent behavior caused by the number of questions. For instance, the large number of choices in Experiment 1 might have led decision makers to switch their decision-making strategy from a compensatory strategy to a simple heuristic of attending only to either reward or time. Thus, in Experiment 2, each respondent saw and made a choice on each option pair only once, not 20 times as in Experiment 1. The drawback is that these data do not permit fine-grained individual level analyses. We interpret the models as describing between-subject heterogeneity and we focus on pooled analyses. Like in the pooled analysis of Experiment 1, it only makes sense to evaluate convex models (that inherently accommodate heterogeneity of preferences wherever multiple core preferences are allowed).

Table 5 gives the model rankings in each stimulus set, according to the pooled Bayes factor, for Experiment 2 (the log transformed Bayes factor values are available in Table S3). Notably, the rankings in this table nearly match those of Experiment 1 in the right panel of Table 4. In particular, the best model in each stimulus set in Experiment 2, according to the PBF, is either noisy-$\mathcal{P}$ or noisy-$\mathcal{PI}$-mix. These models fare well at nearly all τ levels. None of the other models fares particularly well in any stimulus set or with any τ level, with the exception of noisy-$\mathcal{LO}$-mix in the R&R stimulus set.

Insert Table 5 here

To put these results into perspective, recall from Experiment 1 that we found heterogeneity between subjects was best characterized by a mixture of two types of respondents: those attending only to time and those attending only to reward amount (noisy-$\mathcal{P}$ and noisy-$\mathcal{PI}$-mix were the best explanations of the pooled data). If this pattern were merely a consequence of the large number of choices made by each respondent in Experiment 1 then we would expect to see a different pattern in Experiment 2. Since model selection favors the same core in both experiments, we see no reason to suspect a dramatic change. Note that this evidence is only suggestive and not a formal implication, because the aggregate choice proportions do not uniquely identify the mixture components. This is an inherent weakness of analyzing pooled data and the key reason why one can only draw conclusions about individual behavior if one gathers sufficient data from the individual. For instance, choice proportions that are consistent with noisy-$\mathcal{PI}$-mix are also consistent with mixtures of other core theories besides just $\succ_A$ and $\succ_t$. It is possible for noisy-$\mathcal{PI}$-mix to be the best model according to the GBF even when the data are generated by some mixture of compensatory strategies. This problem is particularly vexing for models like noisy-$\mathcal{PI}$-mix, because vectors of choice proportions that are near one-half on every dimension can be generated by nearly limitless combinations of deterministic components. However, in Experiment 2 we actually found that noisy-$\mathcal{P}$ was the best model in four out of six stimulus sets, with $\tau = 0.1$ in one case. The geometry of the parameter space makes it implausible that aggregate data could favor noisy-$\mathcal{P}$ with $\tau = 0.1$ unless the vast majority of individual respondents actually chose according to that model.

8 Roadmap

This paper has been about the interplay between heterogeneity and parsimony in modeling intertemporal preferences. In order to highlight how this issue affects model selection, we have focused specifically on transitive intertemporal preference. Furthermore, instead of considering the menagerie of specific, parametric, transitive theories, we have considered a handful of more general, parameter-free models that are characterized by subsets of viable linear ordered preferences. In particular, we have considered the 'extreme' cases where either just one or two, or all linear orders were considered viable. However, for a given set of stimuli, a parametric theory of the form $u(x) = v(A) \odot \Psi(t)$ typically falls between these two extremes by predicting potentially many, but not all, linear orders as permissible preferences. Other types of theories furthermore predict preferences other than linear orders, such as intransitive preferences. Next, we briefly discuss a roadmap for studying competing theories in a way that formally accounts for heterogeneity. Future analyses of discounting models and intransitive models alike can emulate our approach of modeling either the core preferences, or the responses, or both, as probabilistic processes. Future work can also leverage order-constrained inference methods for statistical inferences and model selection to tackle the complex trade-off between parsimony and heterogeneity. Without much loss of generality and for ease of exposition for rest of this section, we concentrate on the scenario in which two or more theories of the form $u(x) = v(A) \odot \Psi(t)$ compete against each other.

8.1 Stimulus design

Our Stimulus Sets 1-5 are 'standard' intransitivity stimuli in which two attributes trade-off against each other in equal steps as we move through the list of stimuli (similar to the lotteries of Tversky, 1969, in risky choice). Stimulus Set R&R was based on a prior paper on intransitivity of intertemporal preference. If, instead of transitivity, one were rather interested in specific theories of the form $u(x) = v(A) \odot \Psi(t)$, then stimulus design could leverage the specifics of those theories to create choice options that are diagnostic among the theories under consideration. To distinguish these theories, one should use stimuli for which different theories predict minimally overlapping sets of preference patterns. In addition, if the primary goal is to test competing theories (i.e., to either validate or falsify each theory in its own right), one should design the stimuli in such a way that each theory under consideration would also permit as few distinct preference patterns as possible so as to create maximally parsimonious predictions. On the other hand, if the goal is to estimate and identify parameters, say, discount rates, with maximal precision, then one should design stimuli that are maximally diagnostic in that regard, namely, stimuli that lead to many different preference patterns as one varies the discount rate of each theory. In so doing, one ensures that each preference pattern is consistent with only a small range of parameter values of the core theory, say, a narrow range of discount rates. In addition, stimulus design also depends on the type of heterogeneity one wants to either accommodate or critically test.

8.2 Heterogeneity

The type of heterogeneity one wants to account for has strong implications for the type of probabilistic model and level of data aggregation that are suitable. For example, if each individual decision maker satisfies a logit model, but there are individual differences in the parameters of this logit model, then the population generally does not satisfy a logit model because the average of logit probabilities need not be logit probabilities. More generally, if each individual has a core deterministic preference or utility function and only responses are probabilistic, it usually does not make sense to model the population with a single deterministic core preference or utility function, unless it makes sense to treat preferences or utilities as unanimous.

If one were to compare, say, exponential and hyperbolic discounting, it would be advisable to consider multiple different specifications. The first step would be to identify, for the given stimulus set, the set of linear orders that are consistent with exponential and hyperbolic discounting by varying their free parameters. Then, one could consider probabilistic models of the following types.

1. Like our noisy-$\mathcal{P}$, noisy-$\mathcal{I}$, noisy-$\mathcal{PI}$ and noisy-$\mathcal{LO}$ models, it would make sense to consider models with deterministic core preferences that are defined by precisely those linear orders that are consistent with the discounting model at hand, and responses are modeled probabilistically. In addition to the distribution-free error specifications we used, many models of the form $u(x) = v(A) \odot \Psi(t)$, including discounting models, interface naturally with Fechnerian specifications, such as logit and probit models. It is important to reiterate that many of these specifications can be hard to interpret as models of individual behavior if applied exclusively to data pooled across individuals, unless one is willing to assume that those individuals are unanimous in their underlying preferences or utilities.

2. Like our random-$\mathcal{LO}$ and random-$\mathcal{LOT}$ models, it would make sense to consider random preference models that permit a probability distribution over precisely those preference states that are permitted by a given core theory. Because these models feature convex parameters spaces, they can model both within and between person heterogeneity. Interesting parametric special cases to consider, say for exponential and

hyperbolic discounting, are random preference models constructed via a parametric distribution over the permissible discount rates in each core theory.

3. Last but not least, like our noisy-$\mathcal{PI}$-mix and noisy-$\mathcal{LO}$-mix models, it is worth considering hybrid models that permit heterogeneity in the preference states consistent with each given core theory, as well as probabilistic error in responses.

8.3 Model selection criteria

In our analysis, we have emphasized the interplay of heterogeneity and parsimony. In addition to multiple different criteria for goodness-of-fit, we have leveraged the Bayes factor as a model selection tool that is well-suited to quantify parsimony of probabilistic models and to select among models that, like ours, are neither disjoint nor nested. The same methods are useful also for model competitions more generally, including among models based on a core representation of the form $u(x) = v(A) \odot \Psi(t)$. For parametrized theories like that, there are many additional tools available for model selection. For example, some probabilistic models, especially Fechnerian models, naturally plug into adaptive design optimization methods (Cavagnaro et al., 2013) at the individual level. Furthermore, when using models to estimate core parameters, such as an individual's discount rate, it is natural to test the validity of parameter estimates through prediction to new data sets on different stimuli (e.g., the generalization criterion of Busemeyer and Wang, 2000).

8.4 Sketch of a model selection study

We briefly sketch how our roadmap would help design a study aimed at diagnostic design that facilitates replication studies while balancing heterogeneity with parsimony. Table 6 sketches an example of a model competition between exponential and hyperbolic discounting. Imagine that a lab plans a study consisting of a three-stage competition between these two core theories. In Stage I, the lab proposes a set of stimuli that balances two types of diagnosticity: 1) By permitting only few different preference patterns under either theory, it places empirical pressure on both theories. 2) By predicting rather different collections of preference patterns from the two theories, it helps distinguish exponential from hyperbolic discounting. The lab includes several different nonparametric probabilistic models that broadly model probabilistic preferences, or probabilistic responses, or both. The lab also plans frequentist and Bayesian analyses on several different levels, including individual level and group level analyses. In Stage II, the study focuses on the 'best performing' core theory from Stage I to attempt to estimate and identify discount rates. The stimuli for this stage are designed to be maximally diagnostic for that core theory by permitting a broad array of preference patterns as a function of the discount rate. The probabilistic specifications now also include a variety of parametric special cases of the specifications in Stage II. Parametric Fechnerian and random preference models lend additional structure that can help identify discount rates more precisely than the earlier nonparametric models. Depending on the source of heterogeneity, the goal is to obtain either a 'best' single discount rate from each individual or a parametric distribution of each individual's discount rates, or to estimate a population level distribution over discount rates through a variety of probabilistic models and statistical procedures. A major component of Stage II is to evaluate whether and how the 'best' discount rate (point estimate or estimated distribution) varies with the assumed source and the model of heterogeneity. Finally, Stage III is a generalizability study that critically tests the 'best' core theory, 'best' probabilistic specification, and 'best' parameters from Stages I and II on additional stimuli. These stimuli are dependent on the results of Stages I and II and are designed to place maximal pressure on the hypothesized theory, model of heterogeneity, and parameters from Stages I and II. The quantitative performance in all three stages can be evaluated with similar methods.

Insert Table 6 here

9 Conclusions

Heterogeneity causes great challenges in measuring and predicting individual preferences and choices. A common way to think of heterogeneity is that different decision makers might differ in their parameter values (such as their discount rates) within a shared theoretical account (such as exponential discounting) or that a given decision maker might differ in her parameter values for different types of stimuli. Another common way of tackling heterogeneity is to relax restrictions on the functional form of a given theory without changing the probabilistic specification or the response mechanism. Rather than spelling out a refined theory of choice behavior, such approaches pursue increasingly complicated theories of hypothetical constructs. The common practice of inferring parameter values (e.g. discount rates) of a 'prototypical' decision maker from pooled binary choice data of heterogeneous decision makers is rarely grounded in an explicit and compelling model of heterogeneity.

A common way to think of parsimony of a theory is to count the number of parameters in the deterministic core of a theory (and to ear-mark one or more additional parameters for noise or for heterogeneity of parameter values). Counting parameters is only a coarse heuristic in characterizing how flexible or inflexible a theory is in accounting for potential empirical data. As a case in point, on our Set 5, hyperbolic discounting with one free parameter in the algebraic core permits just one preference state, namely $\succ_A$, regardless of the discount parameter. On the other hand, for exponential discounting, which also has one free parameter in the algebraic core, we have found 11 different linear orders, depending on the discount rate. Hence, if we are interested in testing theories empirically, we must keep a close eye on the interplay between the functional form, the probabilistic specification, as well as the stimuli we use in a given study, to account for parsimony in a suitable fashion when analyzing our data. A more rigorous account of model complexity, rather than counting parameters of an algebraic functional form, is to spell out the sources of heterogeneity mathematically and to quantify the flexibility with which the resulting probabilistic model accommodates possible data as a function of the stimuli used.

Here, we aimed to abstract away from distributional assumptions and parametric accounts of heterogeneity and parsimony in intertemporal choice. We focused instead on general characterizations of two crucially important sources of heterogeneity in choices on a given stimulus: the latent intertemporal preferences and the response process. In particular, we considered that the latent preferences may be probabilistic or the responses (based on a given preference) may be probabilistic, or both processes may be probabilistic. While these types of processes have a long history of scientific study, they have been largely neglected in intertemporal choice research. Even though our models differ strongly in their parsimony, every one can be characterized by 10 order-constrained binomial parameters. We have taken a Bayesian approach to quantifying model complexity.

We found that the core preferences $\succ_A$ and $\succ_t$ appeared to drive the performance of the winning models in most cases, suggesting that models draw most of their strength from being able to predict simple patterns of behavior, such as always prefering the highest reward or always prefering the shortest time. However, developing a robust model of intertemporal choice requires attention to a number of issues besides just the core preferences permitted by the underlying theory. Our various levels and types of analyses have shown that both model performance and model selection are sensitive also to the chosen stimulus set, the assumed response process, and whether we analyze data within each individual, jointly across many individuals (GBF), or pooled from many individuals (PBF). We did not find evidence for systematic differences between the U.S and the German study. Also, even though respondents in Experiment 1 each had to handle 20 times as many questions as respondents in Experiment 2, we did not find evidence for systematic differences between the two experiments.

Acknowledgments

This work was supported by the National Science Foundation (SES-1062045 & SES-14-59699 PI: Regenwetter), the Alexander von Humboldt Foundation (TransCoop grant, PI: Stevens), and XSEDE (SES-130016, PI: Regenwetter). We are grateful to Muye Chen for comments on an earlier draft and to Gregor Caregnato for testing respondents at MPI. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of their colleagues, their funding agencies, or their universities.

References

Arfer, K. and Luhmann, C. (2015). The predictive accuracy of intertemporal-choice models. *British Journal of Mathematical and Statistical Psychology*, 68:326–341.

Becker, G., DeGroot, M., and Marschak, J. (1963). Stochastic models of choice behavior. *Behavioral Science*, 8:41–55.

Bernardo, J. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122.

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

Birnbaum, M. (2011). Testing mixture models of transitive preference. Comment on Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, 118:675–683.

Birnbaum, M. and Navarrete, J. (1998). Testing descriptive utility theories: Violations of stochastic dominance and cumulative independence. *Journal of Risk and Uncertainty*, 17:49–78.

Blavatskyy, P. (2011). A model of probabilistic choice satisfying first-order stochastic dominance. *Management Science*, 57:542–548.

Blavatskyy, P. and Pogrebna, G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics*, 25:963–986.

Block, H. and Marschak, J. (1960). Random orderings and stochastic theories of responses. In Olkin, I., Ghurye, S., Hoeffding, H., Madow, W., and Mann, H., editors, *Contributions to Probability and Statistics*, pages 97–132. Stanford University Press, Stanford.

Busemeyer, J. R. and Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, 44(1):171–189.

Carbone, E. and Hey, J. (2000). Which error story is best? *Journal of Risk and Uncertainty*, 20:161–176.

Cavagnaro, D. and Davis-Stober, C. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, 1:102–122.

Cavagnaro, D. R., Gonzalez, R., Myung, J. I., and Pitt, M. A. (2013). Optimal decision stimuli for risky choice experiments: An adaptive approach. *Management science*, 59(2):358–375.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix (Essai on the application of the probabilistic analysis of majority vote decisions)*. Imprimerie Royale, Paris.

Dai, J. (2014). *Using test of intransitivity to compare competing static and dynamic models of intertemporal choice*. PhD thesis, Indiana University.

Dai, J. and Busemeyer, J. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General*, 143:1489–1514.

Davis-Stober, C. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, 53:1–13.

Davis-Stober, C., Brown, N., and Cavagnaro, D. (2015). Individual differences in the algebraic structure of preferences. *Journal of Mathematical Psychology*, 66:70–82.

Doyle, J. (2013). Survey of time preference, delay discounting models. *Judgment and Decision Making*, 8:116–135.

Doyle, J. and Chen, C. (2012). The wages of waiting and simple models of delay discounting. SSRN scholarly paper, Social Science Research Network, Rochester, NY.

Ebert, J. and Prelec, D. (2007). The fragility of time: Time-insensitivity and valuation of the near and far future. *Management Science*, 53:1423–1438.

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

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

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–760.

Green, L. and Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130:769–792.

Guo, Y. and Regenwetter, M. (2014). Quantitative tests of the Perceived Rrelative Argument Model: Commentary on Loomes (2010). *Psychological Review*, 121:696–705.

Harless, D. and Camerer, C. (1994). The predictive value of generalized expected utility theories. *Econometrica*, 62:1251–1289.

Hey, J. (2005). Why we should not be silent about noise. *Experimental Economics*, 8:325–345.

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

Iverson, G. (1990). Probabilistic measurement theory. Available as MBS 90-23 Technical Report at the IMBS, University of California, Irvine.

Iverson, G. and Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, 10:131–153.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Killeen, P. (2009). An additive-utility model of delay discounting. *Psychological Review*, 116:602–619.

Klugkist, I. and Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, 51:6367–6379.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, 112:443–477.

Leland, J. (2002). Similarity judgments and anomalies in intertemporal choice. *Economic Inquiry*, 40:574–581.

Liu, C. and Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52:362–375.

Loewenstein, G. and Prelec, D. (1992). Anomalies in intertemporal choice: evidence and an interpretation. *Quarterly Journal of Economics*, 107:573–597.

Loomes, G., Moffatt, P., and Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty*, 24:103–130.

Loomes, G. and Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, 39:641–648.

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

Luce, R. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, 46:1–26.

Luce, R. (1997). Several unresolved conceptual problems of mathematical psychology. *Journal of Mathematical Psychology*, 41:79–87.

Luce, R. and Narens, L. (1994). Fifteen problems in the representational theory of measurement. In Humphreys, P., editor, *Patrick Suppes: Scientific Philosopher, volume 2: Philosophy of Physics, Theory Structure, Measurement Theory, Philosophy of Language, and Logic*, pages 219–245. Kluwer, Dordrecht.

Luce, R. and Suppes, P. (1965). Preference, utility and subjective probability. In Luce, R., Bush, R., and Galanter, E., editors, *Handbook of Mathematical Psychology*, volume III, pages 249–410. Wiley, New York.

Manski, C. and McFadden, D., editors (1981). *Structural Analysis of Discrete Data with Econometric Applications.* MIT Press, Cambridge, Massachusetts.

Manzini, P. and Mariotti, M. (2006). A vague theory of choice over time. *Advances in Theoretical Economics*, 6:1–27.

Marschak, J. (1960). Binary-choice constraints and random utility indicators. In Arrow, K., Karlin, S., and Suppes, P., editors, *Proceedings of the first Stanford symposium on mathematical methods in the social sciences, 1959*, pages 312–329. Stanford University Press, Stanford, Ca.

Mazur, J. (1984). Tests of an equivalence rule for fixed and variable reinforcer delays. *Journal of Experimental Psychology: Animal Behavior Processes*, 10:426–436.

Mazur, J. (1987). An adjusting procedure for studying delayed reinforcement. In Commons, M., Mazur, J., Nevin, J., and Rachlin, H., editors, *Quantitative Analyses of Behavior: The Effect of Delay and of Intervening Events on Reinforcement Value*, volume 5, pages 55–73. Lawrence Erlbaum Associates, Hillsdale, NJ.

McCausland, W. and Marley, A. (2014). Bayesian inference and model comparison for random choice structures. *Journal of Mathematical Psychology*, 62:33–46.

McClure, S., Ericson, K., Laibson, D., Loewenstein, G., and Cohen, J. (2007). Time discounting for primary rewards. *The Journal of Neuroscience*, 27:5796–5804.

McFadden, D. (2001). Economic choices. *American Economic Review*, 91:351–378.

Myung, J., Karabatsos, G., and Iverson, G. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, 49:205–225.

Pitt, M. and Myung, I. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6:421–425.

Read, D. (2001). Is time-discounting hyperbolic or subadditive? *Journal of Risk and Uncertainty*, 23:5–32.

Read, D. (2004). Intertemporal choice. In Koehler, D. and Harvey, N., editors, *Blackwell Handbook of Judgment and Decision Making*, pages 424–443. Blackwell.

Regenwetter, M., Dana, J., and Davis-Stober, C. (2011). Transitivity of preferences. *Psychological Review*, 118:42–56.

Regenwetter, M., Davis-Stober, C., Lim, S., Guo, Y., Popova, A., Zwilling, C., Cha, Y.-C., and Messner, W. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, 1:2–34.

Regenwetter, M. and Marley, A. (2001). Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*, 45:864–912.

Roelofsma, P. and Read, D. (2000). Intransitive intertemporal choice. *Journal of Behavioral Decision Making*, 13:161–177.

Rubinstein, A. (2003). "Economics and Psychology"? The case of hyperbolic discounting. *International Economic Review*, 44:1207–1216.

Samuelson, P. (1937). A note on measurement of utility. *Review of Economic Studies*, 4:155–161.

Scholten, M. and Read, D. (2006). Discounting by intervals: A generalized model of intertemporal choice. *Management Science*, 52:1424–1436.

Scholten, M. and Read, D. (2010). The psychology of intertemporal tradeoffs. *Psychological Review*, 117:925–944.

Stephan, K., Weiskopf, N., Drysdale, P., Robinson, P., and Friston, K. (2007). Comparing hemodynamic models with DCM. *Neuroimage*, 38:387–401.

Stevens, J. R. (2016). Intertemporal similarity: discounting as a last resort. *Journal of Behavioral Decision Making*, 29:12–24.

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

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34:273–286.

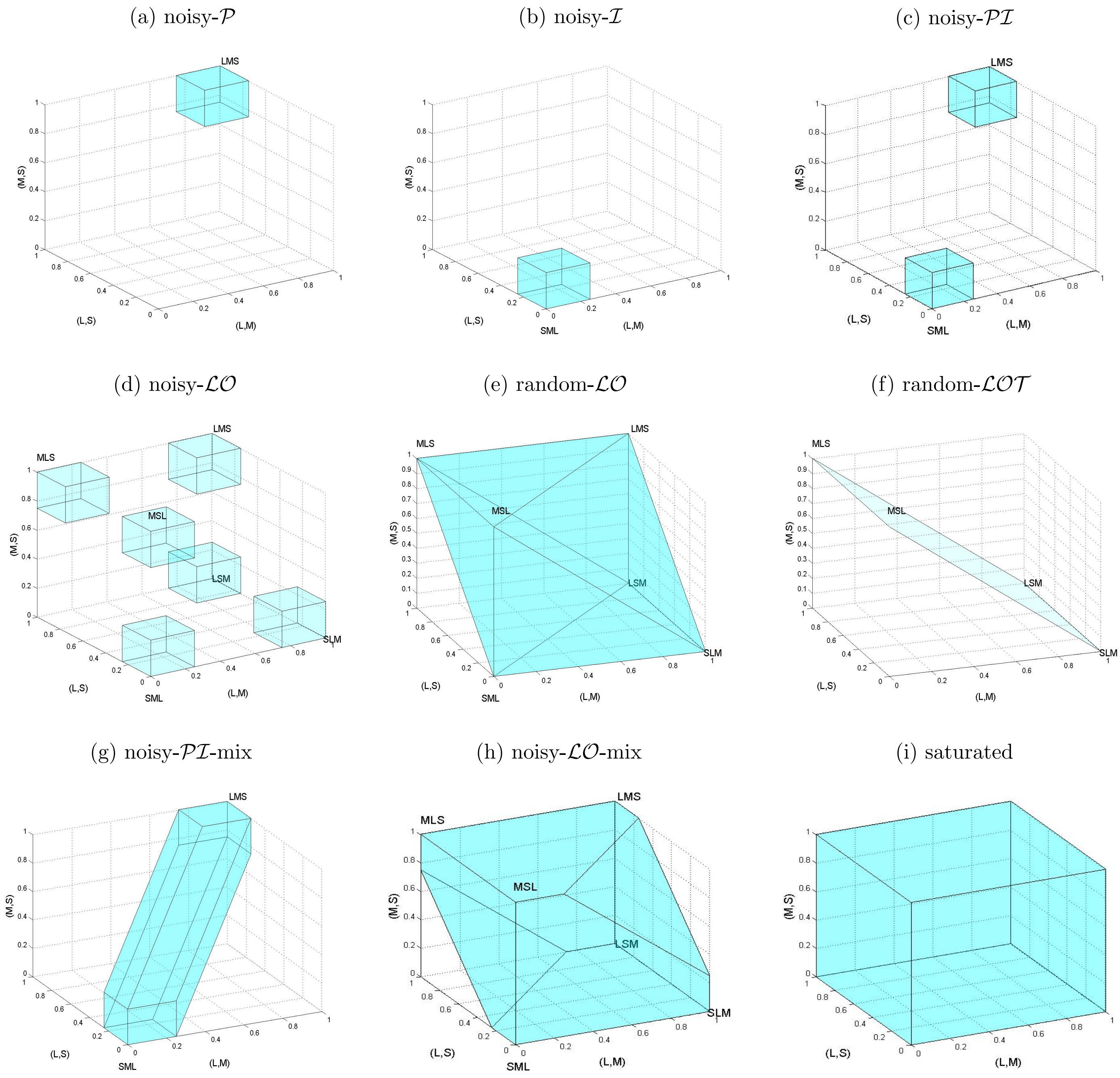
Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G., Roskies, R., Scott, J., and Wilkens-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering*, 16:62–74.

Tsai, R.-C. and Böckenholt, U. (2006). Modelling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology*, 50:1–14.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, 76:31–48.

Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In Cox, J. and Harrison, G., editors, *Risk Aversion in Experiments*, volume 12, pages 197–292. Emerald, Research in Experimental Economics, Bingley, UK.

Yellott, J. (1977). The relationship between Luce's choice axiom, Thurstone's theory of comparative judgement, and the double exponential distribution. *Journal of Mathematical Psychology*, 15:109–144.



(a) noisy-$\mathcal{P}$

(b) noisy-$\mathcal{I}$

(c) noisy-$\mathcal{PI}$

(d) noisy-$\mathcal{LO}$

(e) random-$\mathcal{LO}$

(f) random-$\mathcal{LOT}$

(g) noisy-$\mathcal{PI}$-mix

(h) noisy-$\mathcal{LO}$-mix

(i) saturated

Figure 1: Eight types of probabilistic choice models for linear order intertemporal preferences and the saturated model. The coordinates are the choice probabilities P_{LM}, P_{LS}, P_{MS}. The shaded regions are the permissible choice probabilities for each model. The figure shows the case when $\tau = 0.25$ in (a)-(d); (g)-(h). Considering different upper bounds on error rates yields additional models in these cases.

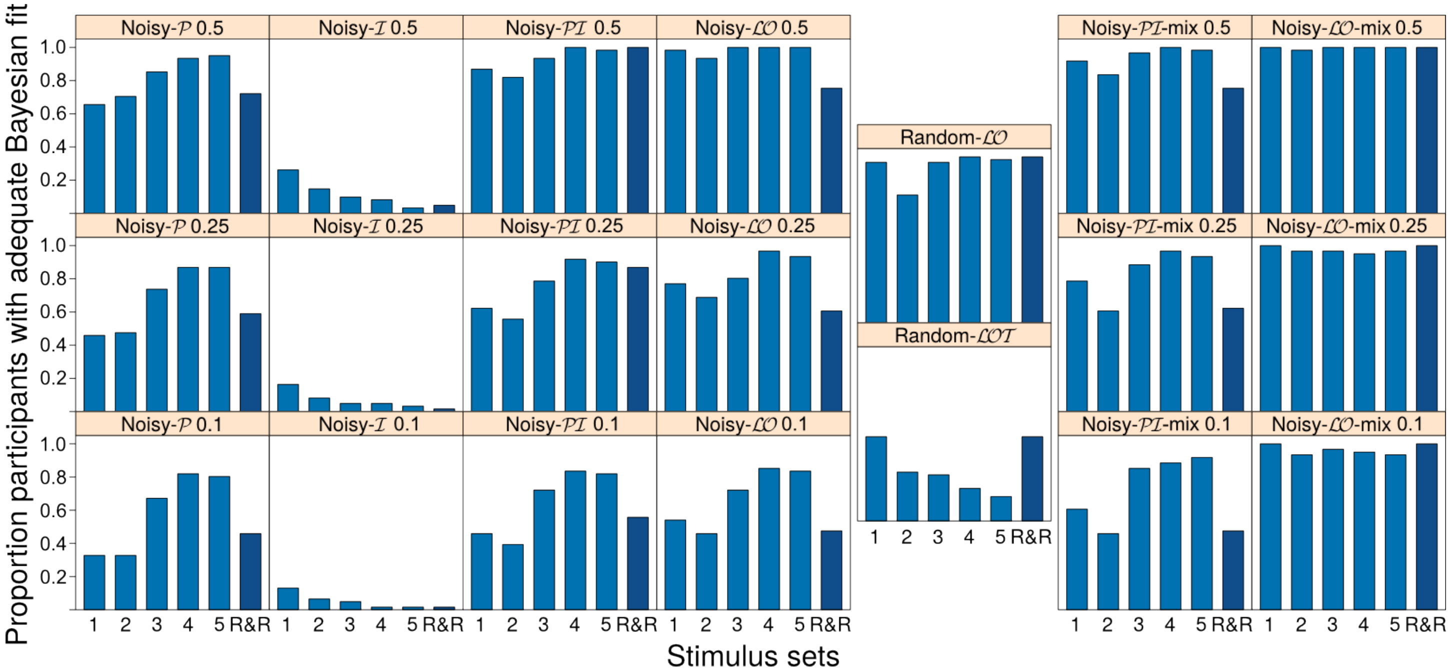


Figure 2: Bayesian p-values in Experiment 1. Each panel shows the results for one model, with the level of τ indicated in the header after the model name (where applicable). Each panel reports the proportion of respondents (out of 61) with adequate fits (Bayesian p-value > 0.05), on the vertical axis, separately for the six stimulus sets.

Table 1: Summary and notational convention for the models under consideration.

Name	**Fig. 1**	**Core Theory**	**Preferences**	**Response Process**	**Label**
noisy-$\mathcal{P}$	(a)	$\{\succ_A\}$	Deterministic	Probabilistic	$\succ_A$
noisy-$\mathcal{I}$	(b)	$\{\succ_t\}$	Deterministic	Probabilistic	$\succ_t$
noisy-$\mathcal{PI}$	(c)	$\{\succ_A, \succ_t\}$	Deterministic	Probabilistic	$\succ_A \vee \succ_t$
noisy-$\mathcal{LO}$	(d)	$\mathcal{LO}$	Deterministic	Probabilistic	$\mathcal{LO}$
random-$\mathcal{LO}$	(e)	$\mathcal{LO}$	Probabilistic	**Deterministic**	$\boldsymbol{\mathcal{LO}}$
random-$\mathcal{LOT}$	(f)	$\mathcal{LO} \setminus \{\succ_{\mathcal{A}}, \succ_{\sqcup}\}$	Probabilistic	**Deterministic**	$\boldsymbol{\mathcal{LOT}}$
noisy-$\mathcal{PI}$-mix	(g)	$\{\succ_A, \succ_t\}$	Probabilistic	Probabilistic	$\succ_A \vee \succ_t$
noisy-$\mathcal{LO}$-mix	(h)	$\mathcal{LO}$	Probabilistic	Probabilistic	$\mathcal{LO}$
saturated	(i)	$\mathcal{B}$	–	–	■

Table 2: Six stimulus sets. In Sets 1-5, we considered all 10 possible distinct *S* vs. *L* pairs among the five listed options. In R&R, we considered the nine listed *S* vs. *L* pairs.

Set 1 options		Set 2 options		Set 3 options		Set 4 options		Set 5 options	
Money	Days	Money	Days	Money	Days	Money	Days	Money	Days
3	4	1	1	14	23	1	1	9	80
5	28	5	21	15	27	3	4	11	83
7	52	9	41	16	31	5	7	13	86
9	76	13	61	17	35	7	10	15	89
11	100	17	81	18	39	9	13	17	92

R&R pairs

S		versus	L	
Money	Days	vs.	Money	Days
7	7	vs.	8	14
7	7	vs.	10	49
8	14	vs.	10	49
10	16	vs.	12	18
10	16	vs.	15	25
12	18	vs.	15	25
4	13	vs.	5	16
4	13	vs.	11	22
5	16	vs.	11	22

Table 3: Experiment 1 - Core theory, preference structure, and response process structure of the best model for each respondent in each stimulus set. Those 31 cases where the best model matches across all six stimulus sets are enclosed in hyphens, such as -2-.

	UIUC Sample							**MPI Sample**					
Respondent	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**R&R**	**Respondent**	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**R&R**
1	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-1-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-2-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-2-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-3-	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	3	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-4-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-4-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-5-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	5	$\mathcal{LO}$	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$
-6-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-6-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
7	$\succ_t$	$\succ_t$	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	7	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
8	$\mathcal{LO}$	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	8	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$
-9-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-9-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
10	$\succ_A \vee \succ_t$	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	-10-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
11	$\succ_t$	$\mathcal{LO}$	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	-11-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
12	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	12	$\succ_t$	$\mathcal{LO}$	$\succ_t$	$\succ_A$	$\succ_A \vee \succ_t$	$\boldsymbol{\mathcal{LO}}$
-13-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-13-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
14	■	■	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	14	$\mathcal{LO}$	■	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$
15	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-15-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
16	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_A$	$\mathcal{LO}$	-16-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
17	$\mathcal{LO}$	■	$\boldsymbol{\mathcal{LOT}}$	$\succ_A$	$\succ_A$	$\boldsymbol{\mathcal{LO}}$	-17-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
18	$\succ_t$	$\mathcal{LO}$	$\mathcal{LO}$	$\succ_A$	$\mathcal{LO}$	$\boldsymbol{\mathcal{LO}}$	18	$\mathcal{LO}$	■	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-19-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-19-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
20	$\succ_t$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	20	$\mathcal{LO}$	■	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-21-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	21	$\succ_t$	$\mathcal{LO}$	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	$\mathcal{LO}$
22	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	22	■	■	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$
23	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	$\succ_t$	$\mathcal{LO}$	-23-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
24	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	-24-	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$	$\succ_A \vee \succ_t$
25	$\succ_A$	$\mathcal{LO}$	$\succ_A \vee \succ_t$	$\succ_A$	$\mathcal{LO}$	$\succ_A$	25	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-26-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-26-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
27	$\mathcal{LO}$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-27-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-28-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-28-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
-29-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	-29-	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$	$\succ_A$
30	$\succ_t$	$\succ_t$	$\succ_A \vee \succ_t$	$\boldsymbol{\mathcal{LOT}}$	$\succ_A$	$\mathcal{LO}$	30	$\succ_A \vee \succ_t$	$\boldsymbol{\mathcal{LOT}}$	$\succ_A \vee \succ_t$	$\succ_A$	$\succ_A$	$\succ_A$
31	$\succ_t$	$\succ_t$	$\succ_A$	$\succ_A$	$\succ_A$	$\mathcal{LO}$	-	-	-	-	-	-	-

Table 4: Experiment 1: Ranking of each model from best to worst, in terms of the Joint (GBF) and Pooled (PBF) analyses, in each stimulus set (column), combined across locations. Rankings in parentheses are worse than the saturated model in the same stimulus set. Ties are given identical ranks. For ease of reading, the three best models, **1**, **2**, and **3**, are marked in boldfaced font.

		Joint (GBF)						**Pooled (PBF)**					
Model	τ	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**R&R**	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**R&R**
noisy-$\mathcal{P}$	0.10	(19)	(18)	(18)	(17)	(16)	(18)	(11)	(9)	(11)	**2**	**1**	(10)
noisy-$\mathcal{P}$	0.25	(17)	(17)	(17)	(16)	9	(16)	(10)	(8)	**2**	**3**	**3**	(9)
noisy-$\mathcal{P}$	0.50	(16)	(15)	(15)	10	10	(13)	**2**	**2**	4	5	5	**1**
noisy-$\mathcal{I}$	0.10	(21)	(21)	(21)	(21)	(21)	(21)	(11)	(9)	(12)	(12)	(12)	(10)
noisy-$\mathcal{I}$	0.25	(20)	(20)	(20)	(20)	(20)	(20)	(11)	(9)	(12)	(12)	(12)	(10)
noisy-$\mathcal{I}$	0.50	(18)	(19)	(19)	(19)	(19)	(19)	(11)	(9)	(12)	(12)	(12)	(10)
noisy-$\mathcal{PI}$	0.10	(15)	(16)	(14)	**2**	**2**	(17)	-	-	-	-	-	-
noisy-$\mathcal{PI}$	0.25	(12)	(13)	**3**	**3**	**3**	(15)	-	-	-	-	-	-
noisy-$\mathcal{PI}$	0.50	**3**	**3**	4	7	7	(9)	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.10	(14)	(12)	(13)	4	5	5	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.25	5	**1**	5	6	6	**1**	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.50	6	**2**	7	9	11	**2**	-	-	-	-	-	-
random-$\mathcal{LO}$		9	(11)	(11)	(14)	(15)	4	4	(7)	6	7	7	4
random-$\mathcal{LOT}$		(13)	(14)	(16)	(18)	(18)	(12)	4	(9)	(12)	(12)	(12)	**3**
noisy-$\mathcal{PI}$-mix	0.10	**1**	(10)	**1**	**1**	**1**	(14)	(11)	(9)	**1**	**1**	**2**	(10)
noisy-$\mathcal{PI}$-mix	0.25	**2**	(6)	**2**	5	4	(11)	**1**	(9)	**3**	4	4	(10)
noisy-$\mathcal{PI}$-mix	0.50	4	(5)	6	8	8	(10)	**3**	**1**	5	6	6	**2**
noisy-$\mathcal{LO}$-mix	0.10	7	(9)	8	11	13	**3**	6	**3**	7	8	8	5
noisy-$\mathcal{LO}$-mix	0.25	8	(7)	9	12	12	6	7	4	8	9	9	6
noisy-$\mathcal{LO}$-mix	0.50	(11)	(8)	(12)	(15)	(17)	(8)	8	5	9	10	10	7
saturated		10	4	10	13	14	7	9	6	10	11	11	8

Table 5: Experiment 2: Ranking of each model from best (highest PBF) to worst (lowest PBF) in each stimulus set. Rankings in parentheses are worse than the saturated model in the same stimulus set. Ties are given identical ranks. For ease of reading, the three best models, **1**, **2**, and **3**, are marked in boldfaced font.

Model	τ	**Set 1**	**Set 2**	**Set 3**	**Set 4**	**Set 5**	**R&R**
noisy-$\mathcal{P}$	0.10	(13)	(12)	(12)	**1**	**1**	(12)
noisy-$\mathcal{P}$	0.25	(11)	4	**3**	**3**	**3**	(10)
noisy-$\mathcal{P}$	0.50	6	**1**	4	5	5	**1**
noisy-$\mathcal{I}$	0.10	(15)	(15)	(15)	(15)	(15)	(15)
noisy-$\mathcal{I}$	0.25	(14)	(14)	(14)	(14)	(14)	(14)
noisy-$\mathcal{I}$	0.50	(12)	(13)	(13)	(13)	(13)	(13)
noisy-$\mathcal{PI}$	0.10	-	-	-	-	-	-
noisy-$\mathcal{PI}$	0.25	-	-	-	-	-	-
noisy-$\mathcal{PI}$	0.50	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.10	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.25	-	-	-	-	-	-
noisy-$\mathcal{LO}$	0.50	-	-	-	-	-	-
random-$\mathcal{LO}$		5	7	6	8	(8)	5
random-$\mathcal{LOT}$		4	(11)	8	(12)	(12)	(8)
noisy-$\mathcal{PI}$-mix	0.10	**2**	5	**1**	**2**	**2**	(11)
noisy-$\mathcal{PI}$-mix	0.25	**1**	**2**	**2**	4	4	(9)
noisy-$\mathcal{PI}$-mix	0.50	**3**	**3**	5	6	6	**2**
noisy-$\mathcal{LO}$-mix	0.10	7	6	7	7	(11)	**3**
noisy-$\mathcal{LO}$-mix	0.25	8	8	9	(10)	(10)	4
noisy-$\mathcal{LO}$-mix	0.50	9	9	10	(11)	(9)	(7)
saturated		10	10	11	9	7	6

Table 6: Sketch of an example model competition between exponential and hyperbolic discounting. A study like this can be pre-registered. It specifies how theories compete, what sources of heterogeneity are permissible, and how they are modeled.

Stage I: Theory testing and screening	
Algebraic Core:	Exponential versus Hyperbolic Discounting
Stimuli:	Permit few preference patterns overall
	Preference patterns diagnostic between these theories
Determ. Pref. & Prob. Resp.	Supermajority specification with three different error bounds
Prob. Pref. & Determ. Resp.	Random preference over permissible preference states
Prob. Pref. & Prob. Resp.	Hybrid model (convex hull of the previous two)
Stage II: Identifying discount rates for best theory from Stage I	
Algebraic Core:	Best theory from Stage I
Stimuli:	Permit many preference patterns
Fixed discount rate & Prob. Resp.	Supermajority specification with three different error bounds
	Logit, probit, Luce, and other Fechnerian models
Prob. discount rate & Determ. Resp.	Parametric random preference over permissible preference states induced by a normal distribution over discount rates
Prob. discount rate & Prob. Resp.	Hybrid models
Stage III: Generalizability to new stimuli	
Algebraic Core:	Same as Stage II
Stimuli:	Permit few patterns based on parameter estimates of Stage II
Model of Heterogeneity:	Best from Stage II

Types of analyses in each stage	
Within subject	frequentist p, Bayes p, Bayes factor
Pooled	frequentist p, Bayes p, Bayes factor
Other	Group Bayes factor, Hierarchical Bayes models