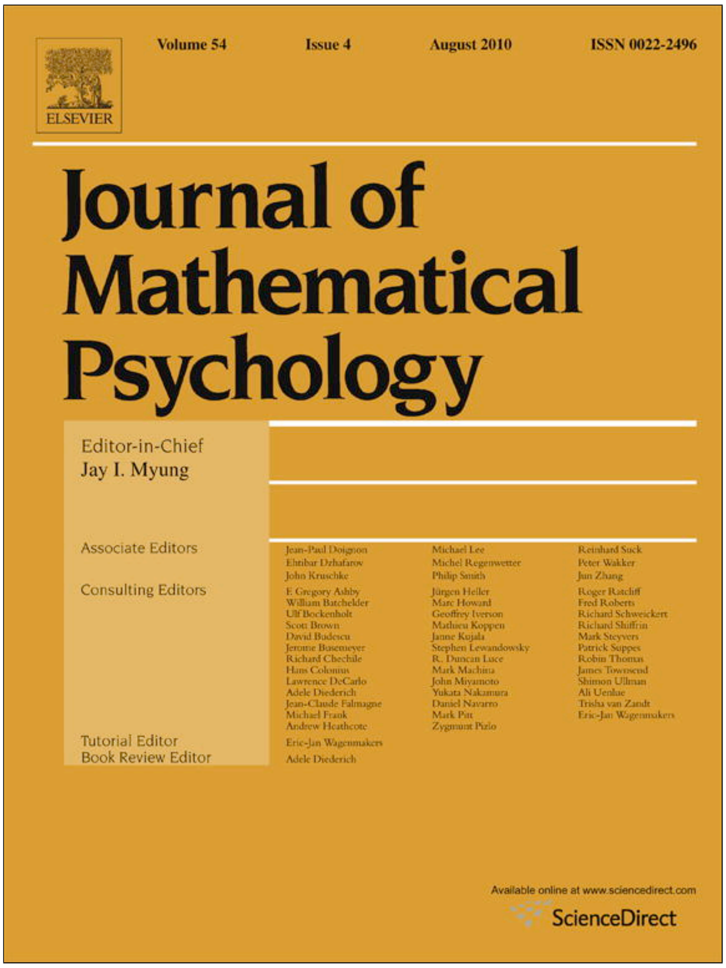
Volume 54
Issue 4
August 2010
ISSN 0022-2496
ELSEVIER
Journal of Mathematical Psychology
Editor-in-Chief
Jay I. Myung
Associate Editors
Consulting Editors
Tutorial Editor
Book Review Editor
Available online at www.sciencedirect.com
ScienceDirect

Journal of Mathematical Psychology 54 (2010) 363–386

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Testing lexicographic semiorders as models of decision making: Priority dominance, integration, interaction, and transitivity

Michael H. Birnbaum

Department of Psychology, California State University, H-830M, P.O. Box 6846, Fullerton, CA 92834-6846, United States

ARTICLE INFO

Article history:
Received 7 February 2007
Received in revised form
17 February 2010
Available online 14 May 2010

ABSTRACT

Three new properties are devised to test a family of lexicographic semiorder models of risky decision making. Lexicographic semiorder models imply priority dominance, the principle that when an attribute with priority determines a choice, no variation of other attributes can overcome that preference. Attribute integration tests whether two changes in attributes that are too small, individually, to be decisive can combine to reverse a preference. Attribute interaction tests whether preference due to a given contrast can be reversed by changing an attribute that is the same in both alternatives. These three properties, combined with the property of transitivity, allow us to compare four classes of models. Four new studies show that priority dominance is systematically violated, that most people integrate attributes, and that most people show interactions between probability and consequences. In addition, very few people show the pattern of intransitivity predicted by the priority heuristic, which is a variant of a lexicographic semiorder model with additional features chosen to reproduce certain previous data. When individual data are analyzed in these three new tests, it is found that few people exhibit data compatible with any of the lexicographic semiorder models. The most frequent patterns of individual data are those implied by Birnbaum's transfer of attention model with parameters used in previous research. These results show that the family of lexicographic semiorders is not a good description of how people make risky decisions.

Risky decision making involves making choices between gambles such as the following:

Would you rather have gamble *A* with 0.5 probability to win \$40 and 0.5 probability to win \$30, or would you prefer gamble *B*, in which you win \$100 with probability 0.5 and win nothing with probability 0.5? Some people prefer *A*, which guarantees at least \$30; whereas other people prefer *B*, which has the higher expected value. Models of risky decision making attempt to describe and to predict such decisions.

The features that differ between gambles such as the probability to win the highest prize, the value of the highest prize, and the value of the lowest prize are called "attributes" of a gamble. We denote gambles *A* and *B* in terms of their attributes as follows: $A = (\$40, 0.5; \$30, 0.5)$ and $B = (\$100, 0.5; \$0, 0.5)$.

The purpose of this paper is to present empirical tests of critical properties that can be used to test models of risky decision making. Critical properties are theorems that can be deduced from one theory but which are systematically violated by at least one rival theory. Three new critical tests are proposed here and evaluated empirically in order to compare a class of lexicographic semiorder models against alternative models. The property of transitivity of preference, which also distinguishes classes of models, is also tested.

Three new properties are described that are implied by lexicographic semiorder models that are violated by other decision making models. *Priority dominance* is the assumption that if a person makes a decision based on a dimension with priority, then no variation of other attributes of the gambles should reverse that decision. *Integrative independence* is the assumption that two changes in attributes that are not strong enough separately to reverse a decision cannot combine to reverse a decision. *Interactive independence* is the assumption that any attribute that is the same in both gambles of a choice can be changed (to another common value) without changing the preference between the gambles. If lexicographic semiorder models are descriptive of how people make decisions, we should not expect violations of these properties, except due to random error. Lexicographic semiorder models can violate transitivity, which is implied by many other models. Let $A \succ B$ represent systematic preference for gamble *A* over gamble *B*. Transitivity is the assumption that if $A \succ B$ and $B \succ C$, then $A \succ C$.

The rest of this paper is organized as follows. The next section describes four classes of decision making models that can be evaluated by testing these four properties. In addition, two specific models are described including parameters chosen from previous research that will be used to make predictions to the new

E-mail address: mbirnbaum@fullerton.edu.

0022-2496/\$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2010.03.002

experiments. The section titled "New Diagnostic Tests" presents the properties to be tested in greater detail. The "Predictions" section presents a summary of the predicted outcomes of the tests according to the classes of models. This section also includes a description of the model of random error that is used to evaluate whether or not a given pattern of observed violations is systematic or can be attributed instead to random variability. The next four sections present experimental tests of priority dominance, attribute integration, attribute interaction, and transitivity, respectively. The results show systematic violations of the three properties implied by lexicographic semiorders that were predicted in advance using a model and parameters taken from previous research. The fourth study searched for violations of transitivity predicted by a specific lexicographic semiorder model; the data failed to confirm predictions of that model. The discussion summarizes the case against lexicographic semiorders as descriptive models of risky decision making, including related findings in the literature. Additional technical details concerning the test of integrative independence, proofs, true and error model, and supplementary statistical results are presented in Appendices.

1. Four classes of decision models

The most popular theories of choice between risky or uncertain alternatives are models that assign a computed value (utility) to each alternative and assume that people prefer (or at least tend to prefer) the alternative with the higher utility (Luce, 2000; Starmer, 2000; Wu, Zhang, & Gonzalez, 2004). These models imply transitivity of preference.

The class of transitive utility models includes Bernoulli's (1738/1954) expected utility (EU), Edwards' (1954) subjectively weighted utility (SWU), Quiggin's (1993) rank-dependent utility (RDU), Luce and Fishburn's (1991; 1995) rank-and sign-dependent utility (RSDU), Tversky and Kahneman's (1992) cumulative prospect theory (CPT), Birnbaum's (1997) rank affected multiplicative weights (RAM), Birnbaum's (1999; Birnbaum & Chavez, 1997) transfer of attention exchange (TAX) model, Marley and Luce's (2001; 2005) gains decomposition utility (GDU), Busemeyer and Townsend's (1993) decision field theory (DFT), and others.

These models can be compared to each other by testing "new paradoxes", which are critical properties that must be satisfied by proper subsets of the models (Birnbaum, 1999, 2004a,b, 2008b; Marley & Luce, 2005). Although these transitive utility models can be tested against each other, they all have in common that there is a single, integrated value or utility for each gamble, that these utilities are compared, and people tend to choose the gamble with the higher utility. For the purpose of this paper, these models are all in the same class.

Let $U(A)$ represent the utility of gamble A. All members of this class of models assume,

$$A \succ B \Leftrightarrow U(A) > U(B), \quad (1)$$

where $\succ$ denotes the preference relation. Aside from "random error," these models imply *transitivity of preference*, because $A \succ B \Leftrightarrow U(A) > U(B)$ and $B \succ C \Leftrightarrow U(B) > U(C) \Rightarrow U(A) > U(C) \Leftrightarrow A \succ C$. That is, because these models represent utilities of gambles with numbers and because numbers are transitive, it follows that preferences are transitive.

Let $A = (x, p; y, 1 - p)$ represent a ranked, two-branch gamble with probability p to win x, and otherwise receive y, where $x > y \geq 0$. This binary gamble can also be written as $A = (x, p; y)$. The two *branches* are probability-consequence pairs that are distinct in the gamble's presentation: (x, p) and $(y, 1 - p)$.

In EU theory, the utility of gamble $A = (x, p; y, 1 - p)$ is given as follows:

$$EU(A) = pu(x) + (1 - p)u(y) \quad (2)$$

where $u(x)$ and $u(y)$ are the utilities of consequences x and y. In expected utility, the weights of the consequences are simply the probabilities of receiving those consequences.

A transitive model that has been shown to be a more accurate description of risky decision making than EU or CPT is Birnbaum's (1999) special transfer of attention exchange (TAX) model. Birnbaum (2008b) has shown that this model correctly predicts data that refute CPT. This model also represents the utility of a gamble as a weighted average of the utilities of the consequences, but weight in this model depends on the probabilities of the branch consequences and ranks of branch consequences in the gamble. In this model, people assign attention to each branch as a function of its probability, but weight is transferred among ranked branches, according to the participant's point of view. A person who is risk-averse may transfer weight from the branch leading to the highest consequence to the branch with the lowest consequence, whereas one who is risk-seeking may transfer attention (and weight) to higher-valued branches. This model can be written for two branch gambles (when $\omega > 0$) as follows:

$$TAX(A) = \frac{au(x) + bu(y)}{a + b} \quad (3)$$

where $a = t(p) - \omega t(p)$, $b = t(q) + \omega t(p)$, and $q = 1 - p$. Intuitively, when the parameter $\omega > 0$ there is a transfer of attention from the branch leading to the best consequence to the branch leading to the worst consequence. This parameter can produce risk aversion even when $u(x) = x$. In the case where $\omega < 0$, weight is transferred from lower-valued branches to higher ones; in this case, $a = t(p) - \omega t(q)$ and $b = t(q) + \omega t(q)$. The formulas for three-branch gambles include weight transfers among all pairs of branches (Birnbaum, 2004a, 2008b; Birnbaum & Navarrete, 1998).

The special TAX model was fit to data of individuals with the assumptions that $t(p) = p^{\gamma}$ and $u(x) = x^{\beta}$; for example, Birnbaum and Navarrete (1998) reported median best-fit parameters as follows: $\beta = 0.41$, $\gamma = 0.79$, and $\omega = 0.32$.

For the purpose of making "prior" predictions for TAX in this article, a still simpler version of the TAX model is used. Let $u(x) = x$ for $0 < x < \$150$; $t(p) = p^{0.7}$, and $\omega = 1/3$. These functions and parameters, which approximate certain group data for gambles with small positive consequences, are called the "prior" parameters, because they have been used in previous studies to predict new data with similar participants, contexts, and procedures. They have had some success predicting results with American undergraduates who choose among gambles with small prizes (e.g. Birnbaum, 2004a, 2008b). Use of such parameters to predict modal patterns of data should not be taken to mean that everyone is assumed to have the same parameters. The main use of these prior parameters is for the purpose of designing new studies that are likely to find violations of rival models.

Several papers have shown that the TAX model is more accurate in predicting choices among risky gambles than CPT or RDU (Birnbaum, 1999, 2004a,b, 2005a,b; Birnbaum, 2006; Birnbaum, 2007; Birnbaum, 2008b; Birnbaum & Chavez, 1997; Birnbaum & McIntosh, 1996; Birnbaum & Navarrete, 1998; Weber, 2007). For example, CPT and RDU models imply first order stochastic dominance, but the TAX model does not. Experiments designed to test stochastic dominance have found violations where they are predicted to occur based on the TAX model with its prior parameters (e.g. Birnbaum, 2004a, 2005a). Similarly, TAX correctly predicted other violations of this class of rank dependent models (Birnbaum, 2004a,b, 2008b). This model has also been extended to make predictions for choice response times as well as choice probabilities and judgments (Birnbaum & Jou, 1990; Johnson & Busemeyer, 2005).

For the purpose of this paper, however, TAX, CPT, RDU, GDU, EU, DFT and other theories in this family are all in the same category and (given selected parameters) can be virtually identical. These models are all transitive, all imply that attributes are integrated, and they all imply interactions between probability and consequences. The experiments in this article will test properties that these integrative models share in common against predictions of other families of models that disagree. Therefore, it should be kept in mind that what is said about TAX in this paper applies as well to other models in its class.

A second class includes *non-integrative but transitive* models. For example, suppose people compared gambles by just one attribute (for example, by comparing their worst consequences). If so, they would satisfy transitivity, but no change in the other attributes could overcome a difference due to the one attribute people use to choose.

A third class of theories assumes that people *integrate* contrasts (differences between the alternatives) but can *violate transitivity* (González-Vallejo, 2002). For the purpose of this paper, this class of theories will be described as *integrative contrast* models, since they involve contrasts and aggregation of the values of these contrasts.

An example of such a model is the *additive contrasts* model:

$$D(G,F) = \sum_{i=1}^{n-1} \phi_i(p_{iG}, p_{iF}) + \sum_{i=1}^{n} \theta_i(x_{iG}, x_{iF}) \qquad (4)$$

where $D(G, F) > 0 \Leftrightarrow G \succ F$, ϕ_i are functions of branch probabilities in the two gambles, and θ_i are functions of consequences on corresponding branches of the two gambles. With gambles defined on positive consequences, these functions are assumed to be strictly increasing in their first arguments, strictly decreasing in their second arguments; they are assumed to be zero when the two components are equal in any given contrast; further, assume

$$\phi_i(p_{iG}, p_{iF}) = -\phi_i(p_{iF}, p_{iG}) \quad \text{and} \quad \theta_i(x_{iG}, x_{iF}) = -\theta_i(x_{iF}, x_{iG}).$$

This type of model can violate transitivity. For example, consider three-branch gambles with equally likely outcomes, $G = (x_{1G}, 1/3; x_{2G}, 1/3; x_{3G}, 1/3)$. Because the probabilities are all equal, all $\phi_i(p_{iG}, p_{iF}) = 0$ in this case. Now suppose

$$\theta_i(x_{iG}, x_{iF}) = \begin{cases} 1 & x_{iG} - x_{iF} > 0 \\ 0, & x_{iG} - x_{iF} = 0 \\ -1 & x_{iG} - x_{iF} < 0 \end{cases} \qquad (5)$$

where $G \succ F \Leftrightarrow D(G, F) > 0$. Let A = (\$80; \$40; \$30) represent a gamble that is equally likely to win \$80, \$40, or \$30. It will be preferred to B = (\$70; \$60; \$20) because $D(A, B) = 1 > 0$ (A has two consequences higher than corresponding consequences of B). Similarly, B will be preferred to C = (\$90; \$50; \$10) for the same reason; however, C will be preferred to A, violating transitivity.

Rubinstein (1988) proposed a model in which small differences are depreciated compared to large differences, and in which people choose the alternative that is better on attributes that are also dissimilar. He showed that such models could account for the Allais paradox. Leland (1994) noted that such models might account for other "anomalies" of choice, which are empirical findings that violate expected utility theory. González-Vallejo (2002) incorporated a normally distributed error component in an additive contrast model called the *stochastic difference model*, and showed that, with specified ϕ_i and θ_i functions, her model could fit several empirical results in risky decision making. These functions incorporated the difference in probability or in consequence divided by the maximum probability or consequence in the choice, respectively.

Integrative contrast models can also violate stochastic dominance. For example, suppose $\phi_i(p_{iG}, p_{iF}) = 100(p_{iG} - p_{iF})$ for $i = 1, 2$; and suppose $\theta_i(x_{iG}, x_{iF}) = x_{iG} - x_{iF}$, for all branches. It follows that G = (\$100, 0.7; \$90, 0.1; \$0, 0.2) will be preferred to F = (\$100, 0.8; \$10, 0.1; \$0, 0.1), even though F dominates G. In this case, $D(G, F) = 100(0.7-0.8)+100(0.1-0.1)+(100-100)+(90-10)+(0-0) = 70$. [To avoid such implications, some have assumed that people first check for stochastic dominance before employing such models (e.g. Leland, 1994). However, systematic violations of stochastic dominance have been observed in empirical studies of such choices (Birnbaum, 1999, 2004a, 2005a, 2008b), which is evidence against CPT, Security Potential/Aspiration model of Lopes and Oden (1999), and other models that satisfy stochastic dominance.]

Another type of contrast model includes products of terms representing probabilities with terms representing contrasts in consequence. A member of this class of interactive models is *regret theory* (Loomes, Starmer, & Sugden, 1991). See the discussion in Leland (1998). Another example is the *most probable winner* model (Blavatskyy, 2006), sometimes called *majority rule* (Zhang, Hsee, & Xiao, 2006). These models can be written as follows:

$$D(G,F) = \sum_{i=1}^{n} \phi_i(E_i)\theta_i(x_{iG}, x_{iF}). \qquad (6)$$

In this case, gambles are defined on mutually exclusive and exhaustive states of the world ("events", E_i), which have subjective probabilities, $\phi_i(E_i)$; and the θ_i are functions that represent, for example, "regret" for having chosen gamble G instead of F given their consequences under that state of the world. This model can also violate transitivity (as in the example of Eq. (5)); but unlike the additive contrasts model (Eq. (4)), this model allows interactions between probabilities of events and their consequences.

A fourth class of models, explored in this paper, includes *non-integrative, non-interactive, intransitive*models. Examples are the lexicographic semiorder (e.g. Tversky, 1969) and the priority heuristic (PH) of Brandstätter, Gigerenzer, and Hertwig (2006) in restricted domains.

A semiorder is a structure in which the indifference relation need not be transitive (Luce, 1956). Suppose two stimuli are indifferent if and only if their difference in utility is less than a critical threshold. Let $\sim$ represent the indifference relation. If $A \sim B$ and $B \sim C$ it need not follow that $A \sim C$.

A *lexicographic order* is illustrated by the task of alphabetizing a list of words. Two words can be ordered if they differ in their first letter. However, if the first letter is the same, one needs to examine the second letters in the two words. At each stage, letters following the one on which they differ have no effect on the ordering of the two words.

In a lexicographic semiorder (LS), people compare the first attribute, and if the difference is less than a threshold, they go on to examine the second attribute; if that attribute differs by less than a threshold, they go on to the third, and so on.

1.1. Lexicographic semiorder model for risky decision making

For analysis of new tests that follow, it is useful to state a LS model for the special case of two-branch gambles with strictly non-negative consequences, $A = (x, p; y, 1 - p)$ and $B = (x', p'; y', 1 - p')$, where $x > y \geq 0$ and $x' > y' \geq 0$. Suppose people consider first the lowest consequences of the two gambles, next the probabilities of the two gambles, and finally, the higher consequences of the two gambles, as follows:

If $u(y) - u(y') \geq \Delta_L$, choose A (7a)

If $u(y') - u(y) \geq \Delta_L$, choose B. (7b)

In these expressions, people choose a gamble if the difference in lowest outcomes exceeds the threshold, Δ_L. Only if neither condition ((7a) or (7b)) holds, does the decider examine the difference between probabilities. In this stage, the decision is made as follows:

$$\text{Else if } s(p) - s(p') \geq \Delta_P, \text{ choose } A \quad (8a)$$

$$\text{Else if } s(p') - s(p) \geq \Delta_P, \text{ choose } B. \quad (8b)$$

Only if none of the above four conditions hold, does the decider compare the largest consequences. In the case of two-branch gambles, the decision would then be based on that attribute alone:

$$\text{Else if } u(x) > u(x') \text{ choose } A \quad (9a)$$

$$\text{Else if } u(x) < u(x') \text{ choose } B. \quad (9b)$$

If none of the above six conditions holds, then the decider has no preference. [1]

In two-branch gambles, there are six possible orders in which to examine the attributes, which will be denoted: *LPH*, *LHP*, *PHL*, *PLH*, *HPL*, and *HLP*. For example, the *LPH* semiorder is the one described in Expressions (7)–(9) in which the lowest consequence, probability, and highest consequence are examined in that order. There are thus four "parameters" in the LS family for two-branch gambles: the first attribute examined, the threshold for deciding on that attribute; the second attribute examined, and the threshold for deciding on that attribute (once the first and second attributes have been chosen, the third is determined for two-branch gambles). In addition, there are two functions, $v(y)$ and $s(p)$ for *LPH*, $u(x)$ and $v(y)$ for *LHP*, etc.

1.2. Priority heuristic

The priority heuristic (PH), as applied to two-branch gambles with strictly positive consequences, is similar to the *LPH* lexicographic semiorder, but it has additional features and assumptions. First, it is assumed that $v(y)$ and $s(p)$ are identity functions. In many applications, such as studies with only two levels of x, p, or y, this assumption adds no constraint. Second, the threshold for consequences is assumed to be one tenth of the largest consequence in either gamble, rounded to the nearest prominent number. [Prominent numbers include integer powers of 10 plus one-half and twice their values (e.g, 1, 2, 5, 10, 20, 50, 100, ...).] This figure yields difference threshold for comparing lowest consequences: $\Delta_L = R[\max(x, x')]$, where R represents the rounded value of one-tenth. For example, if the largest prize of either gamble were \$90, 10% of this would be \$9, which would be rounded to $\Delta_L =$ \$10. (When there are more than two branches in a gamble, it is further assumed that $\Delta_H = \Delta_L$.)

Third, if $|y - y'| \geq \Delta_L$, the decision is made as in Expressions (7a) and (7b). Fourth, if this difference does not reach threshold, then if $|p - p'| \geq \Delta_P = 0.1$, people decide on probability. Fifth, if the difference in probability is not decisive, people are assumed to compare the best consequences. In the case of two-branch gambles, if there is any difference in H, people decide on that factor alone, as in Expressions (9a) and (9b). If none of these reasons is decisive, the priority heuristic holds that people choose randomly between the gambles.

Although the title of Brandstätter et al. (2006) suggests that their model has no "trade-offs," the ratio of a difference to the largest consequence does involve a tradeoff. A difference of \$10 would be decisive in the case of gambles with a maximal prize of \$100, but this difference would not be decisive in a choice with a highest prize of \$1000. So there is a tradeoff between the difference in the lowest consequences and the highest consequence. However, if the largest prize is fixed in a given experiment (e.g., when the largest prize is always \$100), and if there are no more than two branches in a gamble, we can treat the priority heuristic as a special case of the LPH LS model within that restricted experimental domain.

1.3. Comparing models

This article will evaluate the descriptive adequacy of models by two methods. First, the priority heuristic with its threshold parameters is compared against the special TAX model with its prior parameters. To compare models this way, one simply counts the number of people who show the pattern of behavior predicted by each model. This method puts both models on equal footing since no parameters are estimated from the new data. It has the drawback, however, that individual differences or poor selection of parameters could lead to wrong conclusions. In addition, conclusions about the general classes of models are not justified from such tests.

The second method is to test predictions of a model that hold true with any parameters. Testing critical properties has the advantage that it allows rejection of an entire class of models without assuming that all people are the same or that their parameters are known in advance of the experiment. When there are systematic violations, the model whose properties are violated can be rejected, and the model whose predicted violations are observed can be retained. The negative argument (against the disproved class of models) is, of course, stronger than the positive argument in favor of the model that predicted the violations for two reasons: (1) other models might have also predicted the same violations; and (2) some other test of the successful model might lead to its rejection.

Although the second method, testing the critical properties, is the more powerful method in principle, it still depends in practice on parameterized models for appropriate experimental design. We cannot design a very effective test of critical properties unless we use the rival model's parameters to predict where to look for violations. It is easy to conduct a "test" of a property that if violated would be extremely powerful, but in which a class of models and its rivals agree there should be no violation. That is, in principle an experimental design may be fine, but in practice it is not likely to disprove a false model. In order to ensure a good experimental design, therefore, we use the parameterized version of the rival model to design the critical test of the class of models. In this study, we use the parameterized TAX model to design the experiments testing the three properties implied by lexicographic semiorders, and we use the priority heuristic with its parameters to design the tests of transitivity that have potential to unseat the family of transitive models, including TAX.

1.4. Individual differences

Brandstätter et al. (2006) emphasize that the priority heuristic has only one priority order, *LPH* for two-branch gambles. In addition, they assume that there is only one threshold parameter; indeed, they are skeptical about estimating parameters from data.

[1] A reviewer suggested imposing a threshold for the last attribute considered. Introducing such a threshold would imply that the model reaches no decision if all three differences fall short of threshold, even if all three favored the same alternative, contrary to empirical data. Presumably, one would need to add a tie-breaker rule to the LS to construct a more realistic model. Certainly if the tie-breaker included an interactive and integrative component (such as the TAX model), such a hybrid model would be a more realistic description of the data. However, I cannot see why this parameter would improve the accuracy of the LS models without a tie-breaker.

Table 1
Classification of models with respect to testable critical properties. EU = Expected utility theory; CPT = cumulative prospect theory, TAX = transfer of attention exchange model; GDU = gains decomposition utility; OAH = one attribute heuristic; ACM = Additive contrasts model; SDM = Stochastic Difference Model; MPW = most probable winner; RT = regret theory; LS = lexicographic semiorder; PH = priority heuristic.

Model	Testable property			
	Priority dominance	Attribute integration	Attribute interaction	Transitivity
EU, CPT, TAX, GDU	No	Yes	Yes	Yes
OAH	Yes	No	No	Yes
ACM, SDM	No	Yes	No	No
MPW, RT	No	Yes	Yes	No
LS, PH[a]	Yes	No	No	No

[a] The PH allows integration and interaction between lowest and highest consequences when the maximal consequence is varied such that it rounds to different prominent numbers; however, when the highest consequence is fixed, PH agrees with the LS model for two-branch gambles.

They argue that the thresholds are determined by the base-10 number system. Presumably, all people who were educated with the same, base-10 number system should have the same parameters.

However, different people might have different priority orders for the different factors. In addition, within any order, different people might have different threshold parameters for comparing the lowest consequences, highest consequences, and probabilities, Δ_L, Δ_P, and Δ_H.

2. New diagnostic tests

Three new diagnostic tests allow us to test this entire class of LS models against classes of rival models. Whereas LS models have been used in previous studies to account for violations from implications of the family of EU models, in this paper, a member of the family of EU models is used to predict systematic violations of LS models. Relations among theories and the properties to be tested in four new studies are listed in Table 1.

The properties will be stated in terms of three attributes, L, P, and H. These might represent any three factors that can be manipulated, but for the tests presented here, L represents the lowest consequence in a two-branch gamble, P is the probability to win the highest consequence, and H is the highest (best) consequence in the gamble. Let $A = (x_A, p_A; y_A)$ represent the gamble to win x_A with probability p_A and otherwise win y_A, where $x_A > y_A \geq 0$.

2.1. Priority dominance

Priority Dominance holds that attributes with lower priority cannot over-rule a decision based on an attribute with higher priority. Once a critical threshold is reached on an attribute with priority, variation in all other attributes, no matter how great, should have absolutely no effect.

The term "dominance" is used here by analogy to a dominant allele in Mendelian genetics. A heterozygous genotype shows the observed characteristic (phenotype) of the dominant allele. Here, priority dominance refers to the implication that if an attribute has priority and is the reason for one choice, if that same dominant reason is present in a second choice and pitted against improvements of attributes with lower priority, the attribute with higher priority should determine the choice. Hence, no change in "recessive" attributes can overcome a decision based on a dominant attribute.

Priority dominance is the assumption that one and only one of the three attributes has first (highest) priority and that all three of the following conditions hold:

If L has first priority, then for all $y_A > y_B > 0; p_B > p_A > 0; x_B > x_A > 0; p'_B > p'_A > 0; x'_B > x'_A > 0$,

$$A = (x_A, p_A; y_A) \succ B = (x_B, p_B; y_B)$$
$$\Leftrightarrow A' = (x'_A, p'_A; y_A) \succ B' = (x'_B, p'_B; y_B). \quad (10a)$$

If P has first priority, then for all $p_A > p_B > 0; x_B > x_A > 0; y_B > y_A \geq 0; x'_B > x'_A > 0; y'_B > y'_A \geq 0$,

$$A = (x_A, p_A; y_A) \succ B = (x_B, p_B; y_B)$$
$$\Leftrightarrow A' = (x'_A, p_A; y'_A) \succ B' = (x'_B, p_B; y'_B). \quad (10b)$$

If H has first priority, then for all $x_A > x_B > 0; p_B > p_A > 0; y_B > y_A \geq 0; p'_B > p'_A > 0; y'_B > y'_A \geq 0$,

$$A = (x_A, p_A; y_A) \succ B = (x_B, p_B; y_B)$$
$$\Leftrightarrow A' = (x_A, p'_A; y'_A) \succ B' = (x_B, p'_B; y'_B). \quad (10c)$$

Note that in Expression (10a), there is only one reason to prefer $A \succ B$; namely, A has a better lowest consequence. Both of the other attributes favor B. If L has first priority in a lexicographic semiorder, then we know that the difference in lowest consequence was enough to meet or exceed the threshold; otherwise, the person would have chosen B. But if L has first priority, then there should be no effect of any changes in P and H. Therefore, because the same contrast in L is present in A' and B', changing the levels of P and H cannot reverse the preference. Similarly, if $A' \succ B'$, then if L has first priority, then $A \succ B$ by the same argument.

The priority heuristic assumes that the first attribute considered is lowest consequence (L). If L has first priority, and if a \$20 difference is enough to prefer A to B, then no variation in the other two attributes should reverse this preference. For example,

$$A = (\$98, 0.10; \$20) \succ B = (\$100, 0.11; \$0)$$
$$\Leftrightarrow A' = (\$22, 0.1; \$20) \succ B' = (\$100, 0.99; \$0).$$

Suppose, however, that different participants might have different priorities. In that case, we must test all three propositions (that each of the three attributes might be most important). If each person obeys a LS model but different people have different priorities, then each person could show priority dominance in one of the three tests, aside from random error. However, if people use a model with trade-offs such as EU or TAX, then people might be found who violate all three of Expressions (10).

2.2. Attribute integration

Integrative independence is the assumption that small differences on two factors, if not large enough by themselves to reverse a decision, cannot combine to reverse a decision. The term *attribute integration* will be used to indicate systematic violations that cause us to reject integrative independence. This independence property will be tested by factorial manipulations in which two small differences are independently varied. We examine whether the combined effect of changes in two attributes produces a different decision from that produced by each one separately.

Each pair of attributes can be tested for integration by a set of four choices. For example, four choices testing the integration of factors L and H are choices between $A = (x_A, p_A, y_A)$ and

$B = (x_B, p_B, y_B)$, where $x_A > y_A \geq 0$ and $x_B > y_B \geq 0$, can be constructed from a 2×2 factorial design of Contrasts in L [lower consequences, (y_A, y_B) or (y'_A, y'_B)] by Contrasts in H [higher consequences, (x_A, x_B) or (x'_A, x'_B)]. $A = (x_A, p_A, y_A)$ versus $B = (x_B, p_B, y_B)$, $A' = (x'_A, p_A, y_A)$ versus $B' = (x'_B, p_B, y_B)$, $A'' = (x_A, p_A, y'_A)$ versus $B'' = (x_B, p_B, y'_B)$, and $A''' = (x'_A, p_A, y'_A)$ versus $B = (x'_B, p_B, y'_B)$. It is assumed that levels can be selected such that $u(x'_A) - u(x'_B) > u(x_A) - u(x_B)$ and $v(y'_A) - v(y'_B) > v(y_A) - v(y_B)$. Not all choices of levels satisfying the above constraints will produce a diagnostic test between the family of LS models and a particular integrative model.

In order to provide a diagnostic test against a specific integrative model, the ten values in the factorial design above should be chosen such that an integrative model predicts that people will choose B, B', and B'' in the first three choices, and the fourth reverses that decision. It is important to distinguish this pattern, $BB'B''A'''$, which is evidence of integration (both changes are required to produce the change) from the pattern, $BA'A''A'''$ (which is evidence that either manipulation suffices to change the response).

It is possible to find sets of 10 levels for each test such that an integrative model predicts the response pattern $BB'B''A'''$ and no member of the family of LS models can account for this pattern. In this paper, the TAX model with prior parameters is used to design the experiments: levels were selected such that the parameterized TAX model predicts the response pattern $BB'B''A'''$.

Depending on the experimental design, the test may not refute all members of the family of LS models. To assess the diagnostic value of a design, one can work out all possible combinations of parameters to determine if any LS model is compatible with the pattern $BB'B''A'''$. If no version of LS model can predict this pattern and an integrative model predicts it, the experimental design is called a *proper* test of integration. If only a proper subset of the LS models can be refuted, the experimental design is called a *partial* test of integration.

There are many ways to choose ten attribute values to produce a proper test of integrative independence. One method is to choose $x'_A \geq x_A > x_B > x'_B > y'_A > y_A > y_B \geq y'_B \geq 0$, and $p_B > p_A > 0$. Under any monotonic $u(x)$, $v(y)$, and $s(p)$ functions, the above conditions will hold, and the test will be a proper test of integrative independence, in which any LS model implies the following:

$$\text{If } A = (x_A, p_A; y_A) \prec B = (x_B, p_B; y_B), \tag{11a}$$

$$\text{And if } A' = (x'_A, p_A; y_A) \prec B' = (x'_B, p_B; y_B), \tag{11b}$$

$$\text{And if } A'' = (x_A, p_A; y'_A) \prec B'' = (x_B, p_B; y'_B), \tag{11c}$$

$$\text{Then } A''' = (x'_A, p_A; y'_A) \prec B''' = (x'_B, p_B; y'_B). \tag{11d}$$

In this case, note that changing the highest consequences results in the same choice between A' and B' as in the choice between A and B ((11a) and (11b)). Similarly, changing the lowest consequences ((11a) and (11c)) does not reverse the choice between A'' and B'' (it is the same as that between A and B). If people do not integrate information between these attributes, then the combination of both changes (in (11d)) should not reverse the choice between A''' and B'''.

The above constraints imply that there are three mutually exclusive possibilities for the threshold for $H : u(x'_A) - u(x'_B) > u(x_A) - u(x_B) \geq \Delta_H, u(x'_A) - u(x'_B) \geq \Delta_H > u(x_A) - u(x_B)$, or $\Delta_H > u(x'_A) - u(x'_B) > u(x_A) - u(x_B)$. Similarly, there are just three possibilities for the threshold for $L : u(y'_A) - u(y'_B) > u(y_A) - u(y_B) \geq \Delta_H, u(y'_A) - u(y'_B) \geq \Delta_H > u(y_A) - u(y_B)$ and $\Delta_H > u(y'_A) - u(y'_B) > u(y_A) - u(y_B)$. There are two possibilities for P: either $s(p_B) - s(p_A) \geq \Delta_P$ or $\Delta_P > s(p_B) - s(p_A) > 0$. There are therefore $3 \times 3 \times 2 = 18$ threshold patterns by 6 priority orders, yielding 108 LS models to consider. None of these 108 LS models implies the pattern $BB'B''A'''$. When the data show this pattern and none of the possible LS model can predict it, we say that attributes L and H show evidence of *attribute integration*.

Appendix A proves that integrative independence is implied by all LS models in a design satisfying the above constraints. The proof consists of working out all possible combinations of assumptions concerning priority orders and threshold parameters to show that $BB'B''A'''$ is not consistent with any of the LS models. With three attributes (e.g., binary gambles), there are three possible pair-wise tests of attribute integration.

Other tests of integration can be constructed by nesting factorial manipulation of contrasts in different constellations of background levels for the attributes. Because there are ten values to choose, there are many combinations that lead to proper tests of integration. The proofs in each new case consist of working out all possible response patterns produced by the family of LS models that include all possible assumptions concerning the priority order and relations between the thresholds and the contrasts used. Such proofs have been conducted for each of the tests of integration in Tables 4, 6 and 7, as well as for the other properties tested here. Appendix A includes the proof for the test in Table 4, showing that no LS model implies the same response pattern as does the prior TAX model.

Note that if priority dominance held, (11a) is true if and only if (11d) is true, assuming P is highest in priority. Priority dominance must therefore be violated to observe evidence of integration for two "recessive" attributes. However, violation of priority dominance does not rule out integrative independence.

2.3. Attribute interaction

LS models and the priority heuristic assume no interactions between probabilities and consequences. Thus, in a choice between two binary gambles, if all four consequences are held fixed, it should not be possible to reverse the preference by changing the probability to win the larger consequence, as long as that probability is the same in both gambles of each choice. Indeed, if people consider one factor at a time, any factor that is the same in both gambles of a choice should have no effect on the decision. If people are unaffected by any factor that is the same in both alternatives, they will show no attribute interaction. *Interactive independence* for probability and consequences can be defined as follows:

$$A = (x_A, p; y_A) \succ B = (x_B, p; y_B)$$
$$\Leftrightarrow A' = (x_A, p'; y_A) \succ B' = (x_B, p'; y_B). \tag{12}$$

In this case, the probability has been changed but it is the same in both gambles within each choice. In order to find violations, we choose $x_B > x_A \geq 0$, $y_A > y_B \geq 0$, and $p' > p > 0$. For example, $A = (\$55, 0.1; \$20) \succ B = (\$95; 0.1; \$5)$ if and only if $A' = (\$55, 0.9; \$20) \succ B' = (\$95, 0.9; \$5)$ can be refuted. According to any LS model, people should choose either A and A' or B and B', but they should not switch from A to B'. Violations of interactive independence are evidence of *attribute interaction*.[2]

[2] This test of interactive independence should not be confused with the test of interaction in Analysis of Variance, which can be significant even when Expression (12) is satisfied. Nor should it be confused with tests of sign dependence; Expression (12) does not require negative or mixed consequences. Expression (12) has some similarities to but is not the same as Birnbaum's (1974a) "scale-free" test of interaction. Violations of Expression (12) are stronger than Birnbaum's scale-free test because they represent reversals of preference rather than just inequalities of strengths of preference.

A more general form of interactive independence is also implied by LS models and the priority heuristic; namely, as long as the contrast on a given factor is constant,

3. Predictions

The predictions of classes of models are listed in Table 1. The family of integrative, interactive, transitive utility models (including EU, CPT, TAX, and many others) violates priority dominance, violates integrative independence (displays integration), violates interactive independence (shows interaction), and satisfies transitivity. Predictions of the prior TAX model are used to represent this class of models and to show that it can predict violations of the first three properties in each of the tests. (See Appendix B for proofs of properties in Table 1 that are not included in the main text).

One-attribute heuristics (OAH) are transitive and show priority dominance, but they show integrative independence and interactive independence. If a person used only one attribute in making a choice, he or she would not show interactions or integration among attributes.

Additive contrast models, including stochastic difference model (González-Vallejo, 2002), violate priority dominance, exhibit attribute integration, imply no attribute interactions and violate transitivity. Leland's (1994; 1998) version of the Rubinstein (1988) similarity model is a hybrid that uses expected utility as a preliminary step; this means that his version of the similarity model can show integration and interactions between probability and prizes due to its inclusion of EU as a preliminary step.

The majority rule model, most probable winner model, and regret theory (Loomes et al., 1991; Loomes & Sugden, 1982) violate priority dominance; these models show integration and interaction, but they can violate transitivity.

The family of LS models satisfies priority dominance, integrative independence, interactive independence, but it violates transitivity. As long as 10% of the largest consequence rounds to the same prominent number, and we restrict our experiment to two-branch gambles, the priority heuristic is a special case of the LPH LS model. Predictions of LS models will be presented for each of the tests included here.

3.1. Error model

With real data, it is not expected that critical properties will hold perfectly. Indeed, when the same choice is presented to the same person, that person does not always make the same decision. For that reason, we need a model of response variability in order to construct the statistical null hypothesis that a property holds, except for random "error".

For example, suppose we conduct a proper test of integrative independence. According to the LS models, we should not observe response patterns where a person chooses $B \succ A, B' \succ A', B'' \succ A''$ and yet $A''' \succ B'''$. Some violations might arise because a person has the true pattern $BB'B''B'''$, which is compatible with the property, but makes an "error" on the fourth choice. How should one decide that a given number of violations refutes the property?

A "true and error" model allows us to address this question. The error model used here is like that of Sopher and Gigliotti (1993) and Birnbaum (2004b), except with an improvement that uses replications to estimate the error rates. The use of replications avoids the problem that error terms are constructed within the model to be evaluated (Birnbaum & Schmidt, 2008).

When there are replications, error rates can be estimated from cases where the same person makes different decisions when presented the same choices. Consider a choice between a "safe" and a "risky" gamble, denoted S and R, respectively. Let p represent the "true" probability of preferring the safe gamble, S, and let e represent the error rate. If this choice is presented twice, there are four possible response patterns, SS, SR, RS, and RR, where RR denotes choice of the "risky" gamble on both presentations. The theoretical probability that a person would choose the risky gamble on both replicates, $P(RR)$, is then given by the following expression:

$$P(RR) = pe^2 + (1 - p)(1 - e)^2. \quad (13)$$

In other words, this observed pattern can come about in two mutually exclusive ways: people who truly preferred the safe gamble and made two errors or people who truly preferred the risky gamble and twice expressed their preferences correctly. Similarly, the predicted probability of switching from R in the first choice to S in the second choice is given by the following:

$$P(RS) = pe(1 - e) + (1 - p)e(1 - e) = e(1 - e). \quad (14)$$

The probability of making the opposite reversal of preferences, $P(SR)$, is also predicted to be $e(1-e)$. The probability of two choices of the "safe" gamble is $P(SS) = p(1 - e)^2 + (1 - p)e^2$.

This error model has been extended to test such properties as gain loss separability (Birnbaum & Bahra, 2007), integration (Birnbaum & LaCroix, 2008), and transitivity (Birnbaum & Gutierrez, 2007; Birnbaum & Schmidt, 2008); it will be used here to evaluate the properties in Table 1. Additional information is provided in Appendix C.

4. Study 1: A test of priority dominance

According to the class of LS models, once a decisive reason is found, attributes with lower priority are not considered. Therefore, there should be no effect of altering the values of attributes with lower priority, given that there is a decisive difference on the attribute with priority. Priority dominance can be tested in a pair of choices such as the following:

Choice 1: Do you prefer R or S?

R: 10 tickets to win $100
90 tickets to win **$0**

or

S: 10 tickets to win $98
90 tickets to win **$20**

Choice 2: Do you prefer R' or S'?

R': 99 tickets to win $100
01 ticket to win **$0**

or

S': 10 tickets to win $22
90 tickets to win **$20**

In both choices, the risky gamble (R or R') has a lowest consequence of $0 and the safe gamble has a lowest consequence of $20 (bold font). According to the LPH LS model with $\Delta_L = \$10$, people should choose the safe gamble in both cases (S and S', because the lowest consequence has highest priority and is better by $20 in both cases, which exceeds the threshold).

Now suppose that people used a LS model in which the highest consequence had highest priority. If so, people following that priority order might switch from S to R' because the contrast in

there should be no reversal in the effect due to the contrasts on other factors. This more general version, however, requires that we can equate contrasts exactly. In the priority heuristic, for example, it is assumed that equal probability differences are equal (i.e., $s(p) = p$); therefore, the priority heuristic implies this more general definition of interactive independence:

$$A = (x_A, p; y_A) \succ B = (x_B, q; y_B) \Leftrightarrow A' = (x_A, p'; y_A) \succ B' = (x_B, q'; y_B),$$

where $p' - q' = p - q$. This property is not tested in this paper, except in the case where $p = q$.

Table 2
Tests of priority dominance. TAX = predicted response based on prior parameters; "priority order" show predicted preferences under LS models; S = "safe" gamble; R = "risky" gamble. Last three rows show numbers of people (out of 238) showing each pattern.

Choice No.	Choice		TAX	Priority order and threshold parameters									
	Risky gamble (R)	Safe gamble (S)		LPH-1 LHP-1,4	LPH-2 LHP-2,4	LHP-1,3 LHP-1,5	LHP-2,3 LHP-2,5	PLH-1	PLH-2 PHL-3	PHL-4 PHL-5	HPL-3 HLP-2,3	HPL-4 HLP-1,4 HLP-1,5	HPL-5 HLP-2,4
5, 13	10 to win $100 90 to win **$0**	10 to win $98 90 to win **$20**	*S*	***S***	***R***	***S***	***R***[a]	*S*	*R*	*S*	*R*	*S*	*S*[b]
9, 17	99 to win $100 01 to win **$0**	10 to win $22 90 to win **$20**	*R*	***S***	***R***	***S***	***R***	*R*	*R*	*R*	*R*	*R*	*R*
10, 18	10 to win $100 **90** to win $0	90 to win $90 **10** to win $0	*S*	*S*	*S*	*R*	*R*	***S***	***S***	***S***	*R*	*S*	*R*
6, 14	10 to win $100 **90** to win $55	90 to win $60 **10** to win $0	*R*	*R*	*R*	*R*	*R*	***S***	***S***	***S***	*R*	*R*	*R*
7, 19	90 to win **$100** 10 to win $20	90 to win **$60** 10 to win $22	*R*	*R*	*R*	*R*	*R*	*R*	*R*	*R*	***R***	***R***	***R***
11, 15	01 to win **$100** 99 to win $0	99 to win **$60** 01 to win $55	*S*	*S*	*S*	*S*	*S*	*S*	*S*	*S*	***R***	***R***	***R***
8, 20	01 to win **$100** 99 to win $0	01 to win **$22** 99 to win **$20**	*S*	***S***	***R***	***S***	***R***	***S***	***R***	***R***	***R***	***R***	***R***
12, 16	99 to win **$100** 01 to win **$0**	99 to win **$22** 01 to win **$20**	*R*	***S***	***R***	***S***	***R***	***S***	***R***	***R***	***R***	***R***	***R***
Number who showed pattern on first replicate			119	6	2	5	0	0	0	2	1	0	5
Number who showed pattern on second replicate			119	10	1	5	2	0	1	0	1	1	3
Number who showed pattern on both replicates			96	4	0	3	0	0	0	0	0	0	1

Notes: 1. $2 < \Delta_L \leq$ $20; 2. $20 < \Delta_L \leq$ $55; 3. $2 \leq \Delta_H \leq$ $10; 4. $10 < \Delta_H \leq$ $40; 5. $2 < \Delta_H \leq$ $10; all models assume $0 < \Delta_P \leq 0.8$.
[a] LHP-2,5 is same as LHP 2, 3, except it is indecisive on first choice.
[b] HLP-2,4 model is same as HPL-5, except indecisive on first choice.

the highest consequences is too small in the first choice to be decisive ($2), but large enough in the second choice to be decisive ($88). In order to test the hypothesis that highest consequence has priority against the theory that people are integrating information, we need a test in which the highest contrast is fixed and the other two attributes altered. By testing all three hypotheses as to the attribute with highest priority, we can test the whole family of LS models, allowing the possibility that different people have different priorities and different threshold parameters. According to the family of LS models, each person can show priority dominance for one and only one of the three attributes. According to the family of integrative models, priority dominance can be violated for all three attributes.

Table 2 displays three such tests of priority dominance and a fourth test that allow us to evaluate LS models with different priority orders and different ranges for their threshold parameters. The first two choices match those described above, which test the hypothesis that the lowest consequence has priority. The next two choices test if probability has priority, and the following two test if highest consequence has highest priority (bold font). The fourth test (Choices 8, 20, replicated in 12, 16) is a test of interaction. The entries in bold font show the attributes that have fixed values in each test.

With eight choices, there are $2^8 = 256$ possible response patterns in each replicate. The predicted response pattern of the prior TAX model (with parameters used in previous research) are shown in the column labeled "TAX". Each test was devised so that TAX with its prior parameters predicts a reversal of preference in violation of each dominance prediction: the pattern *SRSRRSSR*.

To calculate predictions of the LS models, a spreadsheet of the six priority orders was constructed with each possible combination of parameter values. Given the levels of the design in Table 2, there are four mutually exclusive and exhaustive ranges for Δ_L : $0 < \Delta_L \leq 2, 2 < \Delta_L \leq 20, 20 < \Delta_L \leq 55$, and $55 < \Delta_L$; there are four ranges of Δ_P : $0 < \Delta_P \leq 0.8, 0.8 < \Delta_P \leq 0.89, 0.89 < \Delta_P \leq 0.98, 0.98 < \Delta_P \leq 1$; there are five ranges for Δ_H : $0 < \Delta_H \leq 2$, $2 < \Delta_H \leq 10, 10 < \Delta_H \leq 40, 40 < \Delta_H \leq 78$, and $78 < \Delta_H$. Thus, there are $4 \times 4 \times 5 = 80$ combinations of parameters for each of six priority orders, yielding 480 sets of predictions. There are a total of 33 different response patterns generated by this family of models and parameters, of which 15 patterns had one or more undecided choices. When undecided cases were all resolved as R or all as S, there are a total of 22 distinct response patterns. None of the LS models predicts the pattern, *SRSRRSSR*, predicted by prior TAX.

5. Method of study 1

Participants viewed the materials on computers in the lab connected to the Internet. They made 20 choices between gambles, by clicking a button beside the gamble in each pair that they would rather play. Gambles were represented as urns containing 100 identical tickets with different prize values printed on them. The prize of each gamble was the value printed on a ticket drawn randomly from the chosen urn. Choices were displayed as in the following example:

A: 50 tickets to win $100
   50 tickets to win $0

OR

B: 50 tickets to win $35
   50 tickets to win $25

Table 3
Estimation of "true" and "error" probabilities from replications in tests of priority dominance. TAX shows predictions based on prior parameters. RR = risky gamble chosen on both replicates; RS = risky gamble (R) preferred on the first replicate and safe (S) on the second, etc. Data Patterns show number of people (out of 238) who had each data pattern. Model fit shows estimated true probability of choosing the "safe" gamble ($\hat{p}$), error rate ($\hat{e}$), and test of fit ($\chi^2(1)$); last column shows the $\chi^2(1)$ test of independence of replicates.

No.	Choice		TAX	Data pattern				Model fit			$\chi^2(1)$indep
	Risky Gamble (R)	Safe Gamble (S)		RR	RS	SR	SS	$\hat{p}$	$\hat{e}$	$\chi^2(1)$	
5, 13	10 to win $100 90 to win **$0**	10 to win $98 90 to win **$20**	S	16	11	16	195	0.93	0.06	0.92	54.93
9, 17	99 to win $100 01 to win **$0**	10 to win $22 90 to win **$20**	R	175	18	24	21	0.10	0.10	0.85	37.13
10, 18	10 to win $100 **90** to win $0	90 to win $90 **10** to win $0	S	30	24	18	166	0.86	0.10	0.85	54.33
6, 14	10 to win $100 **90** to win $55	90 to win $60 **10** to win $0	R	176	27	21	14	0.06	0.11	0.75	14.92
7, 19	90 to win **$100** 10 to win $20	90 to win **$60** 10 to win $22	R	195	15	17	11	0.05	0.07	0.12	26.23
11, 15	01 to win **$100** 99 to win $0	99 to win **$60** 01 to win $55	S	14	16	10	198	0.94	0.06	1.37	50.67
8, 20	01 to win **$100** 99 to win $0	01 to win **$22** 99 to win **$20**	S	29	18	22	169	0.86	0.09	0.40	56.42
12, 16	99 to win **$100** 01 to win **$0**	99 to win **$22** 01 to win **$20**	R	183	17	13	25	0.12	0.07	0.53	72.12

They were told that one person per 100 would be selected randomly, one choice would be selected randomly, and winners would receive the prize of their chosen gamble on that choice. This study was included as one among several similar studies on judgment and decision making. Instructions and the materials can be viewed from the following URLs: http://psych.fullerton.edu/mbirnbaum/decisions/prior_dom.htm.

Participants were 238 undergraduates enrolled in lower division psychology courses, 62% were female and 82% were between 18 and 20 years of age.

6. Results of study 1: Violations of priority dominance

The figures in the last three rows of Table 2 show the numbers of people who conformed to each pattern of eight responses in the first replication, second replication, and the number who had the same pattern in both replicates, respectively. The 119 in the first replicate for TAX, for example, indicates that 119 of the 238 participants (50%) had the exact pattern of eight responses predicted by the prior TAX model (with prior parameters) in the first replicate, *SRSRRSSR*. This was the most frequent pattern in the data. The same number showed this pattern in the second replicate. The number in the last row indicates that 96 people showed this exact pattern on all sixteen choices in two replicates.

The last ten columns of Table 2 show predictions of 10 LS models with plausible assumptions about the threshold parameters ($\Delta_P < 0.8, 2 < \Delta_L, 2 < \Delta_H$); these produce 10 sets of predictions, shown in the next ten columns. The model labeled LPH-1 agrees with the priority heuristic of Brandstätter et al. (2006), in which the threshold parameter is $\Delta_L = \$10$. Note that LPH-1 predicts choice of the safe gamble (S) in the first two choices. In contrast, HLP-1, 4 predicts a switch from the safe gamble to the risky one (R) in the first two choices. HLP-1, 4 model assumes that $\$10 < \Delta_H \leq \$40$, and that $2 < \Delta_L \leq \$20$. The listing in Table 2 of models and parameters that make each predicted pattern is not exhaustive (there are 480 model-parameter combinations).

The priority heuristic (LPH-1 in Table 2) was the second most accurate of the 22 prediction patterns of LS models; however, there were only 6, 10, and 4 people whose data were consistent with that model in the first, second, and both replicates, respectively. The most accurate LS model was PHL in which $0.8 < \Delta_P$ and $\Delta_H \leq 10$. This model predicts the pattern, *SRRRRSRR*, which differs from predictions of prior TAX by just two choices, making it the most similar LS model to TAX. This pattern was observed on either the first or second replicate 17 times, including 4 people who repeated it on both replicates. Because this PHL model assumes that $\Delta_P > 0.8$, it would not be regarded as a plausible model (there is ample evidence that people respond to much smaller differences in P).

There were 119 people whose response patterns were the same on both replicates; these displayed only 11 different response patterns; of these, 96 (81%) matched the pattern predicted by prior TAX, only 15 (13%) matched one of the 22 patterns predicted by the 480 LS models; the remaining 8 people had five other patterns. In this case the TAX model with prior parameters outperformed not only the PH with its prior parameters, but it was more accurate than the sum of all of the LS models by a score of 96 to 15.

We can use preference reversals between replicates to estimate "error" rates in the eight choices (Eq. (14)). Table 3 shows how many people made each combination of choices on the two replications of each choice. The parameters, p and e, are estimated by minimizing the $\chi^2(1)$ between predicted and obtained frequencies for the "true and error" model.

The estimated parameters and $\chi^2(1)$ are displayed in three columns of Table 3. The true and error model appears to give a reasonable approximation to the replication data; none of the $\chi^2(1)$ tests are significant. (Critical values of $\chi^2(1)$ are 3.84 and 6.63 for $\alpha = 0.05$ and 0.01, respectively). In contrast, the tests of independence of replications, testing whether the probability of a repeated choice can be represented as the product of two choice probabilities, are all significant, as shown in the last column of Table 3. Both types of tests (true and error versus independence) are based on the same four frequencies, and both estimate exactly two parameters from the data, so these results indicate that the "true and error" model provides a much better description of the data than any theory that implies independence. Expressions (13) and (14) imply independence only when everyone has the same true preferences (i.e., when $p = 0$ or $p = 1$).

If everyone were perfectly consistent with the prior TAX model, for example, the true probabilities of choosing the safe gamble in the first two rows (Choices 5 and 9) should be 1 and 0. Instead,

Table 4
Test of Attribute Integration of L and P (Study 2, $n = 242$). Various Lexicographic Semiorder models make four different patterns of predictions. The TAX model with prior parameters predicts the pattern, *RRRS*.

No.	Choice		% Safe	Predictions of models					
	Risky (R)	*Safe* (S)		TAX	LPH *a*	LHP *b*	PHL, PLH LPH1, *c*	HLP, HPL LHP1	LS Mixture (*Prob choosing* S)
7	90 to win **\$100** 10 to win \$50	50 to win **\$51** 50 to win \$50	11	*R*	*R*	*R*	*R*	*R*	*d*
18	10 to win **\$100** 90 to win \$50	50 to win **\$51** 50 to win \$50	33	*R*	*S*	*R*	*S*	*R*	$a + c + d$
11	90 to win **\$100** 10 to win \$0	50 to win **\$51** 50 to win \$10	22	*R*	*S*	*S*	*R*	*R*	$a + b + d$
8	10 to win **\$100** 90 to win \$0	50 to win **\$51** 50 to win \$10	79	*S*	*S*	*S*	*S*	*R*	$a + b + c + d$
Number of participants who showed the pattern				99	8	21	44	27	

Notes: The priority heuristic agrees with the LPH lexicographic semiorder (LS) model in which $0 < \Delta_L \le \$10$, $\$0 < \Delta_H \le \$49$, and $0 < \Delta_P \le 0.4$. LPH1 and LHP1 are the same LS models as LPH and LHP, respectively, except with $\Delta_L > \$10$. For the LS mixture model, $a =$ the probability of *RSSS*, $b =$ probability of *RRSS*, $c =$ probability of *RSRS*, $d =$ probability of *SSSS*, where $a + b + c + d = 1$. Ten people showed the pattern *SSSS*; no other pattern had more than 8.

the estimated probabilities are 0.93 and 0.10, corrected for the "attenuation" of the error rates, estimated to be $e = 0.06$ and 0.10 for these two choices, respectively.[3]

These results indicate that the LS models cannot be retained as plausible descriptions of individual data. Too many people show a systematic pattern of violation that was predicted in advance of the experiment by a model that violates priority dominance.

7. Study 2: Attribute integration

Brandstätter et al. (2006) criticized two assumptions common to a family of utility models: first, that people summate over values of consequences of mutually exclusive branches in a gamble (integration); second, they doubt that people combine probabilities and consequences by a multiplicative (interactive) relation. Studies 2 and 3 present tests of attribute integration and interaction that allow direct tests of these two assumptions of LS models, respectively.

Table 4 illustrates a test of attribute integration in which the two attributes manipulated are the lower consequences and their probabilities. The highest consequences are fixed to \$100 and \$51 in "risky" and "safe" gambles, respectively. The choices are constructed from a 2 by 2 factorial design of contrast in the lowest consequences (\$50 versus \$50 or \$10 versus \$0) by probability to win in the risky gamble (0.9 versus 0.5 or 0.1 versus 0.5). In the first row (Choice #7), the lowest consequences are the same, but the "risky" gamble has the more favorable probability; therefore the priority heuristic predicts that people should choose *R*. In the second choice (#18), both gambles have the same lower consequence (\$50), but probability decides in favor of the "safe" gamble. Similarly, the priority heuristic implies that the majority should choose the "safe" gamble in the last two choices (#8 and 11) because the lowest consequence of that gamble is \$10 higher than that in the "risky" gamble. This LPH model predicts the pattern *RSSS* for these four choices.

As in Study 1, a spreadsheet was used to calculate predictions for six LS priority orders, combined with parameter assumptions, given the attribute levels selected. In this case, there are $2 \times 2 \times 2 = 8$ parameter combinations to check [\$100 versus \$51 either reaches threshold Δ_H or not; \$10 versus \$0 either reaches threshold Δ_L or not; a 0.4 difference in probability either reaches threshold Δ_P or not]. These 48 (6 priority orders by 8 parameter combinations) yield only 5 distinct patterns: *RRRR*, *RRSS*, *RSRS*, *RSSS*, and *??SS*, where the "?" denotes undecided. Predictions of all 48 LS models are shown in Table 16 of Appendix A. The predicted patterns of LS models are also displayed in Table 4 (the list of models in Table 4 that produce each predicted pattern is not exhaustive). None of the predicted patterns match the predictions of the TAX model with its prior parameters, *RRRS*. Even if we replace *??* with *RR* or *SS*, the pattern, *??SS* does not match this prediction of TAX.

To understand the logic of the design, it helps to start with Choice 7 in the first row of Table 4. In this choice, there is no reason to prefer the safe gamble. In Choice 18, the probability to win the highest prize in the risky gamble has been reduced from 0.90 to only 0.10, which should improve the tendency to choose *S*. But this change is not enough by itself to switch most people to choosing the safe gamble; the majority still chooses the risky gamble. Next, compare Choice 11 with Choice 7; here, the lowest consequences in both gambles have been reduced (from \$50 to \$0 in risky and from \$50 to \$10 in the safe gamble). This manipulation improves the safe gamble, but is not enough by itself to cause the majority to choose the safe gamble. Choice 8 combines both changes. If people do not integrate information, we expect that the combination of two manipulations that separately failed to produce a preference for the "safe" gamble should not combine to produce this preference. Instead, in Choice #8, there is a strong majority (79%) who choose the safe gamble. Two small changes have tipped the scales in favor of the safe gamble.

Note that the pattern *RSSS*, predicted by LPH and the priority heuristic, is not evidence of integration since it represents a decision to choose "safe" if *either* L *or* P favors that gamble; however, the pattern *RRRS* indicates that only when both of these factors favor *S* does the person choose it. In other words, the pattern *RSSS* is consistent with integrative independence (Expressions (11a)–(11d)) whereas *RRRS* violates it.

8. Method of study 2

There were 242 undergraduates from the same pool who viewed the materials in the lab via computers, clicking a button

[3] Suppose everyone had "true" preferences matching those of the TAX model with its prior parameters. Based on the error rates estimated from Table 3, these assumptions imply that $238 \cdot \prod_{i=1}^{8}(1 - e_i) = 118.4$ people should match the predictions of TAX on one replicate and $238 \cdot \prod_{i=1}^{8}(1 - e_i)^2 = 59.0$ should match these predictions on both replicates (to show the pattern perfectly on one replicate, a person must make no errors on eight choices; in order to show the pattern on both replicates, a person must make no errors on sixteen choices). Empirically, Table 2 shows that 119 and 96 showed the exact pattern on one and on both replicates, respectively. This model is clearly an oversimplification, given the evidence of true individual differences (Table 3); however, it is interesting that there are more violations of LS models than expected from this simple model of unreliability, assuming everyone was consistent with prior TAX.

Table 5
Test of Attribute Integration/Interaction, manipulating P and H. The contrast in lowest consequences is fixed. (Study 2, $n = 242$.)

No.	Choice		% Safe	TAX	Predictions of models			
	Risky (R)	Safe (S)			LPH, LHP, PLH *a*	HPL, HLP, PHL *c*	PHL1, LPH1, HPL1, HLP1	LS mixture model
13	10 to win $26 90 to win **$0**	10 to win $25 90 to win **$20**	90	*S*	*S*	*S*	*R*	$a + c$
17	10 to win $100 90 to win **$0**	10 to win $25 90 to win **$20**	75	*S*	*S*	*R*	*R*	a
9	99 to win $26 01 to win **$0**	99 to win $25 01 to win **$20**	71	*S*	*S*	*S*	*R*	$a + c$
4	99 to win $100 01 to win **$0**	99 to win $25 01 to win **$20**	14	*R*	*S*	*R*	*R*	a
	Number of participants who showed the predicted patterns			115	16	25	6	

Notes: LPH, LHP, and PLH assume that $0 < \Delta_L \leq \$20$; HPL, HLP and PHL assume $\$1 < \Delta_H \leq \$75$; PHL1, HLP1, HPL1 assume $0 < \Delta_H \leq \$1$; LPH1 assumes $0 < \Delta_H \leq \$1$ and $\Delta_L > \$20$, LPH2 and LHP2 assume $\Delta_L > \$20$ and $\$1 < \Delta_H \leq \$75$; these last two are indecisive on the first and third choices. According to the mixture model, we can estimate c in two ways: the difference between the first two or last two choice percentages. $\hat{c} = 90 - 75 = 15$, and $\hat{c} = 71 - 14 = 57$, which are quite different.

Table 6
Test of attribute integration, manipulating H and P ($n = 242$).

No.	Choice		% *S*	Predictions of models					LS mixture
	Risky (*R*)	Safe (*S*)		TAX	LPH LHP *a*	LPH1 PHL PLH *b*	HLP LPH2 PHL2 *c*	HPL *d*	
5	10 to win $50 90 to win **$0**	**50 to win $50** **50 to win $49**	91	*S*	*S*	*S*	*S*	*S*	$a+b+c+d$
14	10 to win $100 90 to win **$0**	**50 to win $50** **50 to win $49**	83	*S*	*S*	*S*	*R*	*R*	$a + b$
3	90 to win $50 10 to win **$0**	**50 to win $50** **50 to win $49**	81	*S*	*S*	*R*	*S*	*R*	$a + c$
10	90 to win $100 10 to win **$0**	**50 to win $50** **50 to win $49**	35	*R*	*S*	*R*	*R*	*R*	a
Number of participants who showed each pattern				104	62	25	15	6	

Notes: LPH and LHP assume $0 < \Delta_L \leq \$49$; PHL, PLH assume $0 < \Delta_P \leq 0.4$; LPH1 assumes $\Delta_L > \$49$ and $0 < \Delta_P \leq 0.4$; HPL assumes $0 < \Delta_H \leq \$50$ and $0 < \Delta_P \leq 0.4$; LPH2 assumes $\Delta_L > \$49$ and $\Delta_P > 0.4$; PHL2 assumes $\Delta_P > 0.4$ and $0 < \Delta_H \leq \$50$; HLP assumes $0 < \Delta_H \leq \$50$ and $0 < \Delta_L \leq \$49$.

Table 7
Test of Attribute Integration, manipulating L and H. Probability is fixed, as is the "safe" gamble ($n = 242$).

	Choice		% *S*	TAX	Predictions of models				LS mixture model
	Risky (R)	Safe (S)			LPH LHP PLH *a*	LPH1 LHP1 PLH1 *b*	HLP HPL PHL *c*	HLP1 HPL1 PHL1	
15	**50** to win $51 **50** to win $0	**50 to win $50** **50 to win $40**	94	*S*	*S*	*S*	*S*	*R*	$a + b + c$
1	**50** to win $100 **50** to win $0	**50 to win $50** **50 to win $40**	56	*S*	*S*	*S*	*R*	*R*	$a + b$
6	**50** to win $51 **50** to win $25	**50 to win $50** **50 to win $40**	91	*S*	*S*	*R*	*S*	*R*	$a + c$
12	**50** to win $100 **50** to win $25	**50 to win $50** **50 to win $40**	19	*R*	*S*	*R*	*R*	*R*	a
Number of participants who showed each data pattern				92	25	11	80	2	

LPH, LHP and PLH assume $0 < \Delta_L \leq \$15$; LPH1, PLH1, & LHP1 assume $\$15 < \Delta_L \leq \$40$ and $0 < \Delta_H \leq \$1$; HLP assumes $\$1 < \Delta_H \leq \$50$ and $0 < \Delta_L \leq \$15$; PHL and HPL assume $\$1 < \Delta_H \leq \$50$; HLP1, HPL1, and PHL1 assume $0 < \Delta_H \leq \$1$.

next to the gamble in each choice that they would rather play. Sixteen choices were constructed to form three, 4-choice tests of attribute integration (Tables 4, 6 and 7) and a 4-choice test of interaction and integration (Table 5). There were also three choices containing a test of transitivity (which is described under Study 4). Experimental materials can be viewed at the following URL: http://psych.fullerton.edu/mbirnbaum/decisions/dim_integ_trans2.htm.

9. Results of study 2: Evidence of integration

9.1. Test of Integration of L and P

Examining the percentages in Table 4, the modal choices agree with prior TAX, *RRRS*, since 79% is significantly greater than 50% and 22%, 33% and 11% are all significantly less than 50%. The most frequent pattern of individual data was that predicted by prior TAX, which was exhibited by 99 of the 242 participants. Only 8 people showed the pattern predicted by the priority heuristic, *RSSS*. This comparison (99 to 8) estimates no parameters from either model.

A comparison of the TAX model with free parameters against the family of LS models shows that the LS models cannot account for the *RRRS* pattern whereas TAX with free parameters can handle the other patterns that are compatible with the LS models. Therefore, the predictions of LS are a proper subset of the predictions of TAX. When the true and error model is fit to the data, it can be shown that the TAX model fits the data significantly better than the family of LS models (see Appendix D).

9.2. A mixture model

Suppose people use LS (LS) models, but they randomly switch from one LS model to another (switching both priority order and threshold parameters) from trial to trial. In contrast with the true and error model (which assumes that each person has a single true response pattern), this mixture model would not necessarily show individuals who fit the predicted patterns of individual LS models, because it allows the same person to change priority order and threshold parameters from trial to trial. Because the person is switching LS models from trial to trial, it is possible for a person using this model to exhibit the *RRRS* data pattern. Therefore, this model requires a different statistical test from that described in Appendix D.

We can test whether such a LS mixture model is compatible with the overall choice proportions. Let $a =$ the probability of using a model in which *RSSS* is the pattern, $b =$ the probability of using *RRSS*, $c =$ probability of using a model in which *RSRS* is the pattern, and $d =$ probability of using the pattern *SSSS*. Predictions of this mixture model are shown in the last column, labeled "LS Mixture" in Table 4. Adding the second two choice percentages $(22 + 33 = 55)$, we have $2a + b + c + 2d = 55\%$, which should have exceeded the fourth choice percentage $(a+b+c+d = 79\%)$, if $a > 0$ or $d > 0$. Even if we assume that no one ever used the priority heuristic $(a = 0)$, we still cannot reconcile the observed choice proportions with the theory that people are using a mixture of LS models because 79% is substantially greater than 55%.

The data for four choices consist of eight choice frequencies with four degrees of freedom (within each choice, frequencies sum to the number of participants); the four parameters are restricted to sum to 1, using three degrees of freedom; therefore, one degree of freedom is left to test the LS mixture model. When parameters are selected in the mixture model to minimize the $\chi^2(1)$ comparing the observed choice frequencies with predictions, the minimum $\chi^2(1) = 151.8$, which is significant, as one would expect from the large and systematic deviations between this model's predictions and the data. The TAX model (as fit in Appendix D) yields the following predictions for the choice percentages for Table 4: 12, 33, 22, and 80, respectively, close to the observed values of 11, 33, 22, and 79, respectively.

9.3. Test of Integration/Interaction P and H

Table 5 tests a combination of integration and interaction by manipulating P and H. To test the priority heuristic, the contrast in the lowest consequences was fixed to a larger difference than in Table 4 ($20 versus $0). According to the priority heuristic, the majority should choose the safe gamble in all four rows because the $20 difference in the lowest consequences always exceeds 10% of the highest consequence in either gamble.

The probability to win the higher consequence (in both gambles) and the value of the higher consequence in the risky gamble were varied in a factorial design. In this test, the probabilities are equal in both gambles of a choice, testing interaction.

Instead, the majority choices agree with the prior TAX model's predictions, which was also the most frequent pattern of individual responses (shown by 115 people). Only 16 exhibited the pattern predicted by the priority heuristic with its prior parameters.

Appendix D presents an analysis of the family of LS models allowing all priority orders and assumptions concerning the threshold parameters. The resulting 36 models imply just three response patterns: *RRRR*, *SRSR*, *SSSS*. The TAX model with free parameters can handle these three patterns as well as *SSRR* and *SSSR*. This analysis again refutes the family of LS in favor of TAX with free parameters.

The data of Table 5 allow rejection of the LS mixture model. In this case, the mixture model implies that the choice probabilities in the second and fourth choices in Table 5 (Choices 17 and 4) should be equal. Instead, these are 75% (significantly greater than 50%) and 14% (significantly less than 50%). With best-fit parameters, $\chi^2(2) = 208.0$, so we can reject this mixture model. From the fit of the TAX model, it is possible to calculate predicted choice percentages for Table 5; these are 89, 74, 68, and 16, not far from the obtained values of 90, 75, 71, and 14, respectively.

9.4. Test of Integration with H and P

In Table 6, the "safe" gamble was fixed, and the probability and value of the higher consequence in the risky gamble were varied in a factorial design. In Table 6, a very large difference in the lowest consequence was used in all choices ($0 versus $49). According to the priority heuristic, the majority should choose the safe gamble in every choice because this difference exceeds 10% of the highest consequence. Instead, 91%, 83%, and 81% (all significantly more than half) chose the safe gamble in the first three choices and 65% (significantly more than half) chose the risky gamble in the fourth choice (Choice 10).

The family of 48 LS models with free threshold parameters can handle four possible response patterns: *SRRR*, *SRSR*, *SSRR*, *SSSS*. It can also handle the indecisive pattern, *?R?R*. Resolving both undecided choices in favor of either *S* or *R* results in a total of five patterns, including *RRRR*. In this family it is assumed that either $49 versus $0 reaches threshold or not, that $100 versus $50 either reaches threshold or not, and that a 0.4 difference in probability either reaches threshold or not.

In this test, 104 people showed the response pattern predicted by the TAX model with its prior parameters (*SSSR*); 62 showed the response pattern predicted by the priority heuristic (*SSSS*). When we compare the TAX model with free parameters to the family of LS models, the deviations are again statistically significant (Appendix D).

The LS mixture model also fails to fit the data of Table 6. According to the choice percentages, $\hat{a} = 35$; $\hat{c} = 81 - 35 = 46$; $\hat{b} = 83 - 35 = 48$; therefore, $\hat{d} = 91 - 35 - 46 - 48 = -38$, which is negative, in contradiction to the LS mixture model. With best-fit parameters, $\chi^2(1) = 101.9$, so the LS mixture model can be rejected. In contrast, the TAX model as fit in Appendix D predicts that the four choice percentages for Table 6 should be 92, 84, 81, and 32, very close to the observed values of 91, 83, 81, and 35, respectively.

9.5. Test of integration of L and H

In Table 7, the "safe" gamble was again fixed, all probabilities were fixed to 0.5, and the lower and upper consequences of the risky gamble were manipulated in a factorial design. In this case, the priority heuristic implies that the majority should choose the safe gamble in all four choices. As in other tests, a spreadsheet

Table 8
Probability and prize in the risky gamble were varied in Study 4 ($n = 266$). The "safe" gamble is always the same.

Choice No.	Choice: Risky (R)	Choice: Safe (S)	Series A % S	Series B (reflected) % S	Predictions of models: TAX PHL2	LPH LHP *a*	PLH PHL LPH1 *b*	HLP *c*	HPL *d*	HLP1 HPL1	Mixture model
16	10 to win \$45 90 to win **\$0**	**60 to win \$40** **40 to win \$38**	89	93	*S*	*S*	*S*	*S*	*S*	*R*	$a+b+c+d$
13	10 to win \$100 90 to win **\$0**	**60 to win \$40** **40 to win \$38**	77	81	*S*	*S*	*S*	*R*	*R*	*R*	$a+b$
5	90 to win \$45 10 to win **\$0**	**60 to win \$40** **40 to win \$38**	72	69	*S*	*S*	*R*	*S*	*R*	*R*	$a+c$
9	90 to win \$100 10 to win **\$0**	**60 to win \$40** **40 to win \$38**	27	28	*R*	*S*	*R*	*R*	*R*	*R*	a
	Number of participants showing predicted pattern in series A				93	47	48	26	7	7	
	Number of participants showing predicted pattern in series B				99	45	52	25	2	5	
	Number of participants showing predicted pattern in both A and B				52	26	18	14	0	1	

Series B (Choices 17, 10, 7, and 11) was the same as Series A, except \$100, \$45, \$40, \$38, and \$0 were replaced by \$98, \$43, \$42, \$40, and \$0, respectively, and the positions (first or second) of "risky" and "safe" gambles were reversed. LPH and LHP assume $0 < \Delta_L \leq \$38$; PLH and PHL assume $0 < \Delta_P \leq 0.3$, LPH1 assumes $\Delta_L > \$38$ and $0 < \Delta_P \leq 0.3$; HLP assumes $\$5 < \Delta_H \leq \$60$ and $0 < \Delta_L \leq \$38$, HPL assumes $\$5 < \Delta_H \leq \$60$ and $0 < \Delta_P \leq 0.3$; HLP1 and HPL1 assume $\Delta_H < \$5$. Priority heuristic implies that all choice percentages should exceed 50%. PHL2 makes the same predictions as TAX when $0.3 < \Delta_P \leq 0.5$ and $5 < \Delta_H$.

was constructed to calculate predictions for all six priority orders combined with all parameter ranges that differentiate the choices. In this test, the 36 LS models can produce the patterns, *SSSS*, *SSRR*, *SRSR*, and *RRRR*. In addition, the following patterns with undecided choices are possible: *SS?R*, *SR?R*, *SS??*, *?R?R*, and *????*. The TAX model with prior parameters predicts the pattern *SSSR*. In this test, the *SSSR* pattern might be produced by people with the LHP LS whose "true" pattern is *SS?R*; this design can be considered weaker than the other tests.

The most frequent response pattern by individuals was *SSSR*, which is the pattern predicted by the TAX model with prior parameters.

According to the LS mixture model, $\hat{a} = 19$; $\hat{c} = 91 - 19 = 72$; $\hat{b} = 56 - 19 = 37$, so $\hat{a} + \hat{b} + \hat{c} = 19 + 72 + 37 = 128$, which far exceeds the first choice percentage, 94%. With best-fit parameters, $\chi^2(2) = 338.8$. Thus, these data do not fit the LS Mixture model. According to the TAX model as fit in Appendix D, the choice percentage in Table 7 should be 96, 56, 90, and 19, close to the empirical values of 94, 56, 91, and 19, respectively. Appendix D shows that the family of LS models with fits significantly worse than the TAX model with free parameters.

Although the prior TAX model is the most accurate of the models with fixed parameters, including Table 7, Table 7 is the least convincing of the four tests. Birnbaum and LaCroix (2008) used a replicated test with an improved choice of attribute levels to check the integration of lower and upper consequences. They found stronger evidence of integration of these two attributes.

9.6. A replicated test of P and H

A replicated, factorial test manipulating probability and highest consequence was included in Study 4 (Table 8), which also tested transitivity (discussed under Study 4). The consequences used in Series A and B were so similar that they can be treated as replications. In Table 8, the priority heuristic again predicts that the majority should choose the safe gamble in every choice. Instead, 73% and 72% (significantly more than half) chose the risky gamble in the two replications of Choice 9, consistent with the prior TAX model [with $(\beta, \gamma, \delta) = (1, 0.7, 1)$], which implies the pattern *SSSR*. This pattern (*SSSR*) was shown by 93 and 99 participants in Series A and B, and by 52 people in both series.

TAX (with free parameters) can predict the response patterns, *SSSS*, *SSSR*, *SSRR*, *SRSR*, *SRRR*, and *RRRR*. Parameters that imply these patterns are $(\beta, \gamma, \delta) = (0.5, 0.7, 1), (1, 0.7, 1), (1, 1.5, 0), (2, 0.2, 0), (2, 0.5, -0.5)$, and $(4, 0.2, -1)$, respectively. The test in Table 8 is only a partial test of integration because one LS model can allow *SSSR* (PHL2, with $0.3 < \Delta_P \leq 0.5$ and $5 < \Delta_H$). Appendix E presents an analysis of this replicated test in greater detail.

The mixture model excluding *SSSR* did not fit either Series of Table 8. For Series A, best-fit parameters yielded, $\chi^2(1) = 106.3$; for Series B, $\chi^2(1) = 87.7$. We can therefore reject a mixture model of *SSSS*, *SSRR*, *SRSR*, *SRRR* and *RRRR* (which refutes all LS models except for PHL2 with $0.3 < \Delta_P \leq 0.5$ and $5 < \Delta_H$).

The three tests (Tables 4, 6 and 7) showed that each pair of attributes showed significant evidence of integration. In addition, Table 5 (testing a combination of integration and interaction) is not consistent with any of the LS models. The combination of Tables 4–8 with the study of Birnbaum and LaCroix (2008) allows us to reject the entire family of LS, including the LS mixture model, in favor of the assumption that people integrate each pair of attributes.

10. Study 3: Attribute interaction

The property of attribute interaction is implied by models such as EU, CPT, and TAX. Tests of attribute interaction are shown in Table 9. All four consequences in the choices are the same in all five choices (rows in Table 9). The probabilities change from row to row, but they are always the same within any choice. According to any LS model, including the priority heuristic, there should be no change in preference between any two rows in Table 9.

Any theory that assumes that people make choices by looking strictly at contrasts between the prizes and between probabilities implies no change in preference between the rows. The stochastic difference model (González-Vallejo, 2002) and other additive difference models also assume that people choose by integrating comparisons between values of attributes as well. Although these models are integrative, they assume no interactions between probability and consequences; therefore they also imply interactive independence.

The TAX model and other interactive utility models, in contrast, imply that functions of probability and value multiply. Predictions of the TAX model with prior parameters show that preference should reverse between second and third rows in Table 9.

Table 9
Tests of attribute interaction in Series B of Study 3 ($n = 153$). TAX shows predicted certainty equivalents of risky and safe gambles based on prior parameters; these imply the pattern *SSRRR*.

Choice No.	Choice		% *S*	TAX		Predictions of models	
	Risky (R)	*Safe* (S)		*R*	*S*	LPH, LHP, PLH, PHL, HPL, HLP,	LPH1, LHP1, PLH1, PHL1, HPL1, HLP1
11	01 to win **$95** 99 to win **$5**	01 to win **$55** 99 to win **$20**	83	7.3	20.9	*S*	*R*
17	10 to win **$95** 90 to win **$5**	10 to win **$55** 90 to win **$20**	71	15.6	24.1	*S*	*R*
21	50 to win **$95** 50 to win **$5**	50 to win **$55** 50 to win **$20**	49	35	31.7	*S*	*R*
7	90 to win **$95** 10 to win **$5**	90 to win **$55** 10 to win **$20**	22	54.4	39.2	*S*	*R*
3	99 to win **$95** 01 to win **$5**	99 to win **$55** 01 to win $20	17	62.7	42.4	*S*	*R*

In LPH, LHP, PLH, $0 < \Delta_L \leq \$15$; In LPH1, LHP1, PHL1, $\Delta_L > \$15$. In PHL, HPL, HLP, $\Delta_H > \$40$; in PHL1, HPL1, and HLP1, $0 < \Delta_H \leq \$40$.

Table 10
Frequency of response patterns (Study 3, $n = 153$). Within each repetition, there are 32 possible patterns. The six patterns listed are consistent with interactive utility models; no one repeated any other pattern. Only *SSSSS*, *RRRRR*, and *?????* are consistent with the family of priority heuristic models; *SSSSS* is predicted by the model of Brandstätter et al. (2006). Totals sum to 150 because 3 of 153 participants skipped at least one of the 20 items analyzed here.

Choice Pattern	Series A			Series B		
	Rep 1	Rep 2	Both	Rep 1	Rep 2	Both
SSSSS	10	10	3	14	12	4
SSSSR	3	6	0	7	14	2
SSSRR	11	12	2	39	38	21
SSRRR	50	50	31	35	30	17
SRRRR	18	16	5	14	16	5
RRRRR	22	27	15	15	13	9
Others	36	29	0	26	27	0
Total	150	150	56	150	150	58

11. Method of study 3

Gambles were described either as urns containing 100 tickets, as in earlier studies, or as urns containing exactly 100 marbles of different colors. Trials in the marble format appeared as in the following example:

Which do you choose?

E: 99 red marbles to win $95
01 white marbles to win $5

OR

F: 99 blue marbles to win $55
01 green marbles to win $20.

11.1. Testing attribute interaction

In each variation of the study, there were 21 choices between gambles, with two series of 5 choices testing attribute interaction. Series A consisted of five choices of the form, ($95, p; $5, $1 - p$) vs. ($50, p; $15, $1-p$), in which the five levels of p were 0.01, 0.10, 0.50, 0.90, and 0.99. Series B consisted of five choices of the form, ($95, p; $5, $1-p$) vs. ($55, p; $20, $1-p$), with the same 5 levels of p. The other 11 trials were warm-ups and fillers. Fillers were used so that no two trials from the main design (Series A or B) would appear on successive choices. The two conditions with different formats and fillers gave virtually identical results, so results are combined in the presentation that follows.

Each participant completed 21 choices twice, separated by 5 other tasks that required about 25 min. Materials can be found at the following URL: http://psych.fullerton.edu/mbirnbaum/decisions/dim_x_mhb2.htm

Participants were 153 undergraduates who served as one option toward an assignment in lower division psychology. Of the 153, 78 (51%) were female and 86% were 18–20 years of age.

11.2. Follow-up study: Choice-based certainty equivalents

To further investigate the interaction between probability and prize, and to examine the prominent number rounding assumption in the priority heuristic, a follow-up study was conducted in which participants chose between binary gambles and sure cash. Each choice was of the form, $G = (x, p; \$0, 1-p)$ or certain cash, c. There were 48 choices, constructed from a 4 X 3 X 4, Prize x by Probability p by Certain Cash c factorial design in which Prize $x = \$75, \$100, \$140$, or $200; Probability, $p = 0.1, 0.5$, or 0.9; and Certain Cash, $c = \$10, \$20, \$30$, or $50.

Participants in this follow-up were 231 undergraduates from the same pool. As in earlier studies, gambles were described as containers with 100 equally likely tickets, from which a ticket would be drawn at random to determine the prize. Participants viewed the materials on computers and clicked one of two buttons to indicate preference for the gamble or the cash.

12. Results of study 3: Attribute interactions

The data in Table 9 show a systematic decrease from 83% and 71% choosing the "safe" gamble (both significantly greater than 50%) in the first two rows to only 22% and 17% who do so (both significantly less than 50%) in the last two rows. This systematic decrease violates all LS models, the priority heuristic (people should have chosen *S* in all five choices), the stochastic difference model (González-Vallejo, 2002), other additive difference models, as well as any random mixture of those models. TAX, with prior parameters, correctly predicted the pattern of majority preferences.

Table 10 shows the frequencies of response patterns in both series. There are 32 possible response patterns for each series of five choices, but only the following six were repeated in both replicates by at least one person in at least one series: *SSSSS*, *SSSSR*, *SSSRR*, *SSRRR*, *SRRRR*, and *RRRRR*. Only 3 people agreed with the priority heuristic, *SSSSS*, in both replicates of Series A; 4 did so in Series B, and only 2 of these did so in both A and B. Instead, Table 10 shows that most individuals switched from *S* to *R* in both series as the probability to win the higher consequence was increased, consistent with interactive models.

These same six patterns are all compatible with the family of interactive, integrative models, including TAX, CPT, and EU, among others. The family of LS models can handle only *SSSSS, RRRRR,* and *?????*. Of those who showed the same response pattern in both replicates, 68% and 78% showed patterns compatible with interactive models and not with LS models. Therefore, the family of LS models is not consistent with vast majority of participants in Study 3.

12.1. Follow-up study: Choice-based certainty equivalents

According to the priority heuristic, the cash certainty value in choices between c (sure cash) and $(x, p; 0, 1-p)$, where $0.1 \le p \le 0.9$, depends only on the highest prize in the gamble (x); that is, the choice between gamble and cash depends entirely on x and is independent of p (because all differences in probability exceed threshold). Because 10% of $75, $100, and $140 all round to $10 (the nearest prominent number), there should be no difference between these values of x. People should prefer $10 or more to any gamble with those highest prizes. When the highest prize is $200, people should prefer any cash amount greater than or equal to $20 to the gamble, independent of probability.

Of the 48 choice percentages, the priority heuristic correctly predicted the majority choice in only 20 cases (42% of the choices). For example, 91%, 41%, and 11% preferred $30 over ($100, 0.1; $0, 0.9), ($100, 0.5; $0, 0.5), and ($100, 0.9; $0, 0.1), respectively. According to the priority heuristic, all three percentages should have exceeded 50% since the decision should be based on the difference in lowest prizes ($30), which exceeds 10% of the highest prize. Similarly, 93%, 41%, and 15% preferred $50 to ($200, 0.1; $0, 0.9), ($200, 0.5; $0, 0.5), and ($200, 0.9; $0, 0.1), respectively. Such reversals of the majority preference related to probability occurred for all sixteen combinations of highest prize and certain cash, contrary to the priority heuristic.

In order to solve for the cash equivalent value of each gamble (the value of c that would be judged higher than the gamble 50% of the time), choice proportions were fit to the following interpolation equation:

$$P_{ijk} = \frac{1}{1 + \exp[a_{ij}(g_{ij} - c_k)]} \qquad (15)$$

where P_{ijk} is the predicted probability of choosing cash value c_k over gamble $(x_i, p_j; \$0, 1-p_j)$; the values of c_k were set to their cash values ($10, $20, $30, $50), and the logistic spread parameter a_{ij} was estimated separately for each gamble. The values of g_{ij} are the certainty equivalents of the gambles; that is, the interpolated cash values that would be judged better than the gamble 50% of the time. These were estimated from the data to minimize the sum of squared differences between the predicted probabilities and observed choice proportions.

The interpolated certainty equivalents of the gambles are shown in Fig. 1 as a function of the largest prize value (x), with a separate curve for each level of probability to win, p. According to the priority heuristic, all three curves should coincide. That is, the cash-equivalent value should depend only on x in these choices and should be equal to Δ_L. Instead, the curves never coincide.

According to the priority heuristic, gambles with highest prizes of $75, $100, and $140 should all show the same cash equivalent values, since people should have rounded 10% of these values to $10, the nearest prominent number. Therefore, the curves in Fig. 1 should have been horizontal until $140 and only show a positive slope between $140 and $200. This might be the case when probability is 0.1 (lowest curve), but the curves show positive slope when the probability to win is either 0.5 or 0.9. Instead of showing the patterns predicted by the priority heuristic, the curves diverge to the right, consistent with a multiplicative interaction between prize and probability, such as is assumed in utility models like EU, CPT, TAX, and others.

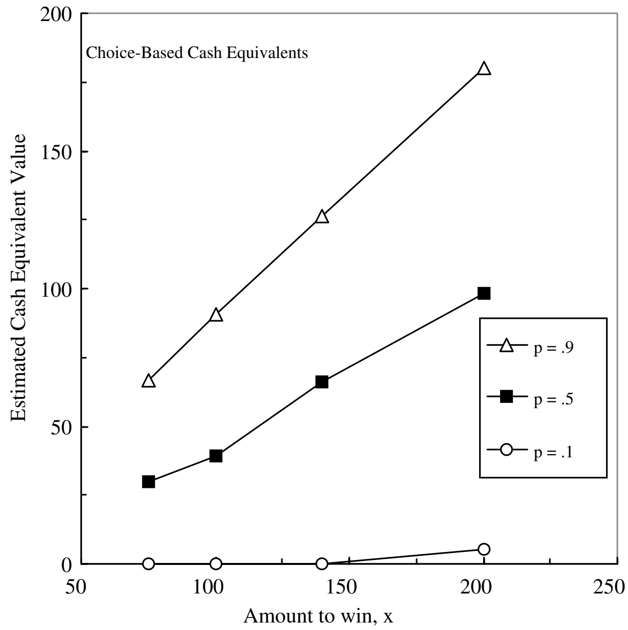

Fig. 1. Cash equivalent values of binary gambles, win x with probability p, otherwise $0, plotted as a function of x with a separate curve for each p. According to the priority heuristic, all three curves should coincide, because probability should have no effect. In addition, the (single) curve should be horizontal for the first three cash values, since they all round to the same prominent number, $100. Instead, the curves show a divergent, bilinear interaction, consistent with a multiplicative combination of functions of probability and prize.

13. Study 4: Transitivity of preference

In Studies 1–3, tests of priority dominance, attribute integration, and attribute interaction all show that the family of LS models can be rejected because data systematically violate predictions of those models. In those tests, LS models had to defend the null hypothesis against integrative and interactive models that predicted violations.

It should be clear that tests of integration and interaction can lead to only three outcomes; they can retain both families of models, they could reject LS and retain the integrative, interactive family, or they might reject both families. There is no way for the LS models to "win" over the integrative, interactive models in these tests because the LS models can handle only a subset of the response patterns consistent with integrative, interactive models in those studies. However, it is possible to refute the family of transitive models (including TAX, CPT, EU) and retain the family of LS models in tests of transitivity (which is tested in Study 4). Testing transitivity, therefore, allows LS models such as the priority heuristic to go on offense, putting transitive models like TAX in danger of refutation.

The first three choices in Table 11 provide a new test of transitivity in which the priority heuristic predicts violations. Similarly, the second set of three choices (Series B) is a near replicate in which the priority heuristic also implies violation of transitivity, with the counterbalance that positions of the gambles are reversed.

Methods matched those of the earlier studies. Participants were 266 undergraduates from the same pool, tested in the lab, who completed the experiment at the following URL: http://psych.fullerton.edu/mbirnbaum/decisions/dim_int_trans4.htm.

There were 18 choices between gambles consisting of 4 warm-ups, 8 trials forming a replicated test of factorial manipulation of prize and probability in the risky gamble (Table 8), and 6 trials composing two tests of transitivity. Choices were presented in restricted random order such that no two choices testing transitivity would appear on successive trials.

Table 11
Test of transitivity in Study 4 ($n = 266$). According to the priority heuristic, majority choices should violate transitivity.

Choice No.	Choices in series C		% Second	Models	
	First (*F*)	Second (*S*)		TAX	LPH
6	70 to win $100 30 to win $0	78 to win $50 22 to win $0	16	*F*	*F*
18	78 to win $50 22 to win $0	85 to win $25 15 to win $0	14	*F*	*F*
14	85 to win $25 15 to win $0	70 to win $100 30 to win $0	84	*S*	*F*

Priority heuristic assumes $\Delta_P = 0.1$. Series D, Choices 15, 12, and 8 were the same as Series C (#6, 18, and 14, respectively), except $100, $50, and $25 were replaced with $98, $52, and $26, respectively, and the positions of the gambles (first or second) were counterbalanced. The corresponding (reflected) choice percentages are 20%, 19%, and 86%, respectively.

Table 12
Analysis of response patterns in test of transitivity in Study 4. The TAX model with prior parameters implies the pattern, *FFS*.

Choice pattern			Frequency of choice patterns		
#6 15	#18 12	#14 8	Series C	Series D (reflected)	Both
F	*F*	*F*	18	9	0
F	*F*	*S*	184	177	149
F	*S*	*F*	10	11	1
F	*S*	*S*	12	17	0
S	*F*	*F*	8	8	1
S	*F*	*S*	18	22	2
S	*S*	*F*	7	9	3
S	*S*	*S*	9	13	2
Column totals			266	266	158

14. Results of study 4: Transitivity

The priority heuristic predicts that the majority should choose the second gamble in Choices 6 and 18; the lowest prizes are both $0 and the difference in probability is .08 or less, so people should choose by the highest consequence. In these two choices, it makes the correct prediction. However, in Choice 14, PH predicts that people should choose the first gamble, which has a lower probability to receive $0 (0.15 rather than 0.30, a difference exceeding 0.10). Instead, 84% chose the gamble with the higher prize. Similar results were observed in the second series. So, the majority data fail to show the predicted pattern of intransitivity predicted by the priority heuristic.

Table 12 shows the number of people who showed each response pattern in the two tests. The most frequent response pattern (Choosing the first gamble (*F*) in the first two choices and the second gamble (*S*) in the third, *FFS*) is displayed by 184 and 177 of 266 participants in Series C and D (reflected), respectively, and by 149 people (56%) in both tests. This pattern matches the predictions of the TAX model with prior parameters (it is also consistent with other transitive utility models). The intransitive pattern predicted by the priority heuristic, *FFF*, was shown by 18 and 9 people in Series C and D, respectively, but no one showed that pattern on both series. The family of LS can handle *FFF*, *FFS*, *FSS*, *SFF*, *SSF*, *SSS*, *S?S*, *S?F*, *F?S*, *F?F*, *??S*, *??F*, and *???*; Note that it can accommodate both intransitive patterns, *FFF* and *SSS*, either of which would violate the transitive models like TAX.

When the data of Table 12 were fit to the true and error model (Appendix F), the best-fit estimates of the model indicated that no one was truly intransitive.

Study 2 ($n = 242$) also included a test of transitivity in binary gambles in which the lowest consequence was $0. The choices are shown in Table 13. The majority preferences were transitive. The modal pattern was to prefer *A* over *B*, *B* over *C* and *A* over *C*. However, 41 people (17%) showed the intransitive pattern predicted by the priority heuristic (*FFF*). Fitting the "true and error" model to the response frequencies, an acceptable fit was obtained for a solution with only three "true" response patterns (all transitive): *FFS*, *FSF*, and *SSF*, with estimated "true" probabilities of 0.86, 0.05, and 0.09, respectively, with error terms of 0.10, 0.02, and 0.24, respectively. Deviations from this purely transitive model were not significant, $\chi^2(1) = 1.75$.

Had people been systematically intransitive in the manner predicted by the priority heuristic, then all of the transitive utility models such as TAX, RAM, CPT, GDU, EU, and others would have been rejected. However, empirical tests found no systematic evidence of intransitivity where predicted by the priority heuristic.

15. Discussion

All four studies yield clear results that either violate or fail to confirm predictions of LS models and the priority heuristic. When we compare models with prior parameters, very few participants show the data patterns predicted by the priority heuristic, whereas the TAX model with prior parameters did a good job of predicting the most common data patterns in these studies. When we compare classes of models by testing agreement with critical properties, we find that evidence against the family of LS models including the priority heuristic is very strong and evidence against the family of transitive models is very weak or nonexistent.

Tests of priority dominance in Study 1 show that we can reject the hypothesis that most people show priority dominance for one attribute. Instead, the evidence indicates that if there are people who show this property, their number must be very small.

Tests of attribute integration show that each pair of attributes shows evidence of integration for a large and significant number of participants. Small changes that do not reverse a decision by themselves can combine to reverse the modal preference. The theory that people use a mixture of LS models with different priority orders and different threshold parameters, (switching from trial to trial) does not account for the choice proportions in any of the four tests in Study 2. The least impressive case against LS models in Study 2 was the test of integration of lowest and highest consequences (Table 7); however, Birnbaum and LaCroix (2008) found strong evidence against the LS models with an improved test of this property.

Tests of attribute interaction in Study 3 show clear evidence of interaction between probability to win and amount to win. Study

Table 13
Test of transitivity ($n = 242$) included in study 2.

Choice no.	First (*F*)	Second (*S*)	TAX	Priority heuristic	% S
2	A: 60 to win $60 40 to win $0	B: 67 to win $40 33 to win $0	*F*	*F*	17
16	B: 67 to win $40 33 to win $0	C: 73 to win $20 27 to win $0	*F*	*F*	16
19	C: 73 to win $20 27 to win $0	A: 60 to win $60 40 to win $0	*S*	*F*	70

3 shows that preferences can be reversed by changing the value of probability that is the same in both gambles in a choice. LS models and additive difference models require that any attribute that is the same in both choices cannot affect the choice. Instead, the data show very clear violations. Although Brandstätter et al. (2006) argued that humans do not have addition or multiplication in their "adaptive toolbox," there is strong evidence against their arguments in Studies 2 and 3. Further, the follow-up to Study 3 (Fig. 1) found that people behave as if the certainty equivalents of gambles involve multiplication between functions of probability and of prize value.

When individual differences are analyzed in these new tests, it is found that few participants produce response patterns compatible with the LS models, including the priority heuristic. The most common pattern of individual data is that predicted by the TAX model with its prior parameters, not that of any of the LS models.

Tests of transitivity show that few, if any, participants exhibit systematic violations of transitivity of the type predicted by the priority heuristic and other LS models. Failure to find intransitivity does not prove there is none to be found, of course. Design of a study of transitivity will be sensitive to assumed threshold parameters and the priority order of the LS model used to devise the test. These studies were designed using the priority heuristic with its assumed threshold parameters.

Although some previous studies reported systematic violations of transitivity based on the Tversky (1969) gambles (see reviews in González-Vallejo, 2002, Iverson & Myung, 2010), others who have reviewed this literature concluded that the evidence does not warrant abandonment of transitive theories (Iverson & Falmagne, 1985; Luce, 2000; Regenwetter, Dana, & Davis-Stober, in press; Rieskamp, Busemeyer, & Mellers, 2006; Stevenson, Busemeyer, & Naylor, 1991). Interestingly, Tversky went on to publish transitive theories (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992).

Birnbaum and Gutierrez (2007) tested transitivity using the same gambles as in Tversky (1969) and using gambles with prizes 100 times as large. The LS used by Tversky (1969) and the priority heuristic of Brandstätter et al. (2006) predict that the majority of participants should show intransitivity with these gambles. For example, most should prefer $A = (\$500, 0.29; \$0)$ over $C = (\$450, 0.38; \$0)$, C over $E = (\$400, 0.46; \$0)$, and the majority should violate transitivity by choosing E over A. The highest estimated rates of intransitivity observed by Birnbaum and Gutierrez occurred in a condition of their second experiment in which they used pie charts to display probability and no numerical information on probability was provided to participants. In this condition, with either 160 undergraduates tested in the lab or 201 Web recruits tested via the WWW, the estimated percentage who were consistent with the intransitive strategy, was 5.8% in the true and error model. Out of 361 participants, only 14 (3.9%) displayed the intransitive pattern predicted by the priority heuristic on two repetitions. Although greater than zero, an empirical incidence of 3.9% and theoretical incidence of 5.8% agreement with the predicted data pattern seems quite low for a theory that is supposed to be descriptive.

Birnbaum and LaCroix (2008) also tested transitivity with 50–50 gambles: $A = (\$100, 0.5; \$20)$, $B = (\$60, 0.5; \$27)$, and $C = (\$45, 0.5; \$34)$. People should choose $A \succ B$ and $B \succ C$, but $C \succ A$, according to the priority heuristic (which assumes $\Delta_L =$ \$10. In this case, only one person out of 260 tested showed the pattern of intransitive choices predicted by the priority heuristic and one showed the opposite pattern.

In an extensive review of the literature and reanalysis using a random utility model, Regenwetter et al. (in press) concluded that a transitive, random utility model can be retained for the vast majority of participants tested in the literature. A distinction can be drawn between the random utility model and the true and error model. In the true and error model, each subject is assumed to have a single "true" pattern of preferences that may or may not be transitive. The transitive random utility model is a mixture model in which each subject is assumed to randomly choose one of the many possible transitive orders on each trial and make each decision based on that randomly chosen order. Unfortunately, many of the data sets preserved from early publications (e.g., data of Tversky, 1969) have not been saved in a form that allows reanalysis via the true and error model.

15.1. Evaluation of the priority heuristic

Few people satisfied the critical properties of a LS model. Still fewer appeared consistent with the priority heuristic. Using prior parameters for both models, TAX was more accurate than PH. In Study 1 (Table 2), there were 96 people whose response patterns matched the prior TAX model on both replicates against just 4 whose data matched the predictions of PH on both replicates. In Study 2 (Tables 4–7), the figures are 99, 115, 104, and 92 for prior TAX against 8, 16, 62, and 25 for PH, respectively. In Table 8, there were 52 people matching TAX on both replicates against 26 for PH. In Study 3 (Table 10) there were 31 and 17 people who matched prior TAX on both replicates against 3 and 4 who matched the PH. Finally, in Study 4 (Table 12) there were 184 and 177 people whose data matched TAX compared with 18 and 9 for PH.

Given how poorly the PH did in predicting these results, how could that model have had such seemingly "good" accuracy in fitting certain previously published data, as claimed by Brandstätter et al. (2006)? Two points should be made. First, they did not attempt to fit certain previously published data (e.g. Birnbaum & Navarrete, 1998) in which their heuristic was not accurate. Second, their analyses relied on a global index of fit (percentage of correct predictions of the majority choice) and they did not estimate parameters from the data. Comparisons of fit can easily lead to wrong conclusions when parameters are not properly estimated (Birnbaum, 1973, 1974b). When parameters are estimated for all three models from the data, CPT, TAX, and the PH all fit the Kahneman and Tversky (1979) perfectly. Therefore, those data should be considered not diagnostic. When parameters were estimated for other data sets analyzed by Brandstätter, et al., both CPT and TAX either matched or outperformed the priority heuristic with its parameters also estimated (Birnbaum, 2008a). So, their conclusion that the PH is an accurate model was not justified even for the data analyzed by Brandstätter et al. (2006).

15.2. Expected value exclusion

Brandstätter et al. (2006, p. 426) noticed that their priority heuristic was not accurate in cases where expected value (EV) differed by a ratio greater than 2. Suppose people first compute the EV of each gamble; take their ratio, choose the gamble with the higher EV if the ratio exceeds 2, and use the PH otherwise. (It is doubtful that Brandstätter, et al. would endorse this model, since they argue against the kinds of assumptions that it seems to entail and it contradicts their "frugality" argument that people do not use all the information). The EV model implies attribute integration and interaction, violates priority dominance, and satisfies transitivity as long as the threshold for using EV is small enough. The manipulations in Studies 1, 2, and 4 fall outside this EV exclusion zone in many instances. This EV-modified priority heuristic model could therefore handle some of the results reported here, but EV does not account for all of the violations of the priority heuristic either in this paper or in previous research.

For example, in Table 9, the priority heuristic predicts that people should choose the "Safe" gamble in all choices, including

Choices 17 and 7. In these choices, EV ratios are less than 2), as they are in the last four choices of Table 9). But we see a clear effect of the common probability in these choices, contrary to the PH.

Similarly, the PH predicts that people should prefer the "safe" gamble in Choice 18 of Table 4 (it has a lower probability of the lowest consequence). Instead, 67% choose the "risky" gamble, ($100, 0.1; $50, 0.9) over the "safe" gamble, ($51, 0.5; $50, 0.5) despite the fact that the EV ratio is only 1.09.

Violations of stochastic dominance reported by Birnbaum (1999, 2004a,b, 2005a), Birnbaum (2006), Birnbaum (2008b) represent violations of EV as well as violations of the PH. Violations of restricted branch independence and cumulative independence (Birnbaum & Navarrete, 1998) also include cases where people violate both expected value and the PH. For example, $R = (\$97, 0.1; \$11, 0.1; \$2, 0.8)$ has a higher EV (12.4) than $S = (\$49, 0.1; \$45, 0.1; \$2, 0.8)$, whose EV is 11; however, most people choose S, contrary to both EV and the priority heuristic. Brandstätter, Gigerenzer, and Hertwig (2008) replicated a subset of such choices used by Birnbaum and Navarrete (1998) and found that PH predicted fewer than half of the modal choices in a new sample of Austrians, despite the fact that EV ratios were close to 1.

Glöckner and Betsch (2008) devised tests that distinguish CPT and PH, given their prior parameters, while controlling for EV. The priority heuristic made systematic errors that were correctly predicted by CPT.

Birnbaum (2008c) devised a new critical test that compares the TAX model to both CPT and the priority heuristic, while controlling EV. Choices were constructed from the following recipe:

$$R = (x, p - r; x^-, r; z, 1 - p) \quad \text{or} \quad S = (y, q; z^+, s; z, 1 - q - s)$$

and

$$R' = (x, p; z', r'; z, 1 - p - r')$$
$$\text{or} \quad S' = (y, q - s'; y^-, s'; z, 1 - q)$$

where $x > x^- > y > y^- > z^+, z' > z \geq 0$. According to CPT with any parameters and functions, $R \succ S \Rightarrow R' \succ S'$. (Note that R' dominates R and S dominates S' by first order stochastic dominance.) Because the lowest consequences are the same in both gambles, the priority heuristic predicts that people will choose $S \succ R$ when the probability of the lowest consequence $(1 - q - s)$ is lower than the corresponding probability in the risky gamble $(1 - p)$ and this difference is greater than or equal to the threshold (0.1). For example, people should choose $S = (\$66, 0.8; \$8, 0.1; \$7, 0.1)$ over $R = (\$92, 0.6; \$90, 0.1; \$7, 0.3)$; and they should choose $R' = (\$92, 0.7; \$8, 0.2; \$7, 0.1)$ over $S' = (\$66, 0.7; \$63, 0.1; \$7, 0.2)$. Instead, the opposite pattern, RS', was significantly more frequent than the pattern compatible with either CPT or the PH. In sum, even the addition of EV as a preliminary step in the priority heuristic would not rescue the PH from failed predictions.

16. Summary and conclusions

Some investigators have argued that successes of algebraic models in psychology are only "as if" approximations of what people are "really" thinking when they make decisions. These investigators might argue that people really reason by processes akin to language rather than to perception. That is, decision processes might follow discrete, binary logic, rather than analog numerical functions. These studies show that it is possible to test these notions when they are specified and they find no evidence to support this view; instead, the data appear best represented by models such as EU, CPT, and TAX that assume that people act as if they form a weighted average of the utilities of the consequences and choose the gamble with the higher weighted average, apart from random error. [Perhaps "as if" should be applied to all theories that have thus far been proposed or to none, since this term has been poorly defined in a way that fails to differentiate theories in the literature.]

This study illustrates testing critical properties to compare theories. If one theory is true, the critical property follows; therefore, if the property is violated, we can reject the theory. If a rival theory is true, we can use it with its parameters to design a test that will compare the parametric predictions of the rival theory against the theorem of the other. In this case, the TAX model with prior parameters was used to design the tests implied by the family of LS. Unless the rival model is used to design the experiment, it would be easy to assemble an experiment such that both theories could account for the results. Similarly, the PH was used to design studies of transitivity.

A difference between testing critical properties and "fitting" models to data is illustrated in the analysis of Study 1. A reviewer expressed surprise that the TAX model with prior parameters could be more accurate than the sum of all 480 LS models constructed by allowing six different priority orders and 80 combinations of parameters. Of the 256 possible response patterns, one is predicted by prior TAX and 22 are consistent with various members of the LS models. How could one beat the sum of 22?

The reason for the reviewer's surprise, I think, is that in post-hoc fitting exercises in which data collected for some other purpose are fit to rival models, models with more parameters tend to achieve better fits than models with fewer. They have more degrees of freedom to capitalize on chance. However, in tests of critical properties, this principle need not hold. Critical properties are properties that hold for any specification of functions and parameters, so the apparent "freedom" given to a model in such tests is illusory, not real.

One reviewer of this paper argued that testing critical properties is not equally "fair" to both models. This reviewer blamed the perceived injustice on the statistical tests, but the "unfairness" is strictly due to the logic of science and not to statistics. In the first three studies, the tests had the potential to refute the family of LS models in favor of TAX, but no result would have resulted in refutation of TAX in favor of LS because the implications of the LS family are a subset of those of TAX. However, TAX could predict outcomes that were inconsistent with the family of LS models. The only result that might have favored LS in those studies was if the violations in Studies 1–3 were nonsignificant, in which case both models could be retained (though the prior parameters would need revision). The TAX model was in jeopardy only in Study 4, where violations of transitivity could have unseated it, along with other transitive models.

But this "unfairness" of science does not mean that results testing critical properties are a foregone conclusion. Had systematic violations of transitivity predicted by PH been observed in Study 4 and had the properties of priority dominance, integrative independence, and interactive independence been sustained in Studies 1–3, the conclusions of this paper would be the opposite of what they are.

Reviewing the findings with respect to Table 1, we see that not only the family of LS models but also one attribute heuristics can be rejected by the strong evidence showing violation of priority dominance, violation of integrative independence and violation of interactive independence.

Additive difference models, including the stochastic difference model violate priority dominance and are compatible with attribute integration, so these models are compatible with the results of Studies 1 and 2, but they assume no interaction, so they are thus not consistent with the results of Study 3.

Regret theory and majority rule are interactive as well as integrative and can therefore handle the first three studies. Although this class of models violates transitivity, it can handle

Table 14
A test of attribute integration. To devise a test, one selects levels intended to make G preferred in the first choice, and to select levels of $u(a'_F)-u(a'_G) > u(a_F)-u(a_G)$ and $v(b'_F)-v(b'_G) > v(b_F)-v(b_G)$ so that F''' will be preferred in the fourth choice. The choice pattern, *GGGF*, violates integrative independence, which refutes all possible lexicographic semiorder models, as shown in Table 15.

Choice	Gamble, G	Gamble, F
1	$G = (a_G, b_G, c_G)$	$F = (a_F, b_F, c_F)$
2	$G' = (a'_G, b_G, c_G)$	$F' = (a'_F, b_F, c_F)$
3	$G'' = (a_G, b'_G, c_G)$	$F'' = (a_F, b'_F, c_F)$
4	$G''' = (a'_G, b'_G, c_G)$	$F''' = (a'_F, b'_F, c_F)$

the results of Study 4. Birnbaum and Schmidt (2008) tested for specific violations of transitivity predicted by regret theory and majority rule; they found few exceptions to the conclusion that everyone was transitive. Nevertheless, this class of intransitive models remains compatible with the main findings in the present paper (see Appendix B).

The integrative, transitive utility models (like EU, CPT, and TAX) are compatible with these results because they are consistent with the violations of critical properties (they were predicted by the TAX model with prior parameters) and because the data do not yet require us to reject transitivity of preference.

The weakest part of this argument is the conclusion that transitivity is acceptable. Finding conformity to a principle does not imply that there are no violations to be found. Therefore, the focus for empirical investigations for proponents of models like the regret model, majority rule, priority heuristic, or stochastic difference model appears to be to demonstrate where consistent and systematic violations of transitivity can be observed. Negative findings in this paper and other recent studies shift the burden of proof to supporters of intransitive models: show us where to find evidence of systematic intransitivity.

Acknowledgments

This research was supported by grants from the National Science Foundation, SES-0202448 and SES DRMS-0721126. Thanks are due to Jeffrey P. Bahra, Roman Gutierrez, Adam R. LaCroix, Cari Welsh, Christopher Barrett, and Phu Luu, who assisted with pilot studies and data collection for this research. Thanks are due to Eduard Brandstätter for providing a preprint of his paper in 2005 and for discussions of these issues, and to Jerome Busemeyer, Anthony J. Bishara, Daniel Navarro Martinez, Jörg Rieskamp and two anonymous reviewers for comments on an earlier draft.

Appendix A. Attribute integration and lexicographic semiorders

Suppose three attributes (A, B, and C) can vary in each alternative. The ABC LS model assumes that people examine these attributes in the order: First A, then B, then C, using the following rule to decide between two alternatives, $G = (a_G, b_G, c_G)$ and $F = (a_F, b_F, c_F)$:

If $u(a_G) - u(a_F) \geq \Delta_A$ choose G
Else if $u(a_F) - u(a_G) \geq \Delta_A$ choose F
Else if $v(b_G) - v(b_F) \geq \Delta_B$ choose G
Else if $v(b_F) - v(b_G) \geq \Delta_B$ choose F
Else if $t(c_G) - t(c_F) > 0$ choose G
Else if $t(c_F) - t(c_G) > 0$ choose F
Else undecided

where Δ_A, Δ_B, and Δ_C are difference threshold parameters for comparison of attributes A, B, and C, respectively, and $u(a)$, $v(b)$, and $t(c)$ are strictly increasing, monotonic functions of the attribute values. LS models generated by other priority orders, ACB, BAC, BCA, CAB, and CBA are similarly defined.

A test of integration of attributes A and B (with levels of C fixed) is illustrated in Table 14. It is assumed that levels can be selected such that $u(a'_F) - u(a'_G) > u(a_F) - u(a_G)$ and $v(b'_F) - v(b'_G) > v(b_F) - v(b_G)$. Note that the levels in this test are selected so that G is favored in the first choice and both manipulations tend to improve F relative to G.

Integrative independence of A and B is the property that

If $G = (a_G, b_G, c_G) \succ F = (a_F, b_F, c_F)$
And $G' = (a'_G, b_G, c_G) \succ F' = (a'_F, b_F, c_F)$
And $G'' = (a_G, b'_G, c_G) \succ F'' = (a_F, b'_F, c_F)$
Then $G''' = (a'_G, b'_G, c_G) \succ F''' = (a'_F, b'_F, c_F)$.

If people integrate attributes A and B, then integrative independence can be violated with the pattern $G \succ F$, $G' \succ F'$, $G'' \succ F''$ and $F''' \succ G'''$, which is denoted, *GGGF*. It can be shown that with appropriate selection of the ten attribute levels, none of the LS models imply this response pattern. Table 15 presents the predicted response patterns of all possible combinations of priority orders together with all possible assumptions concerning whether the contrasts in each attribute reach threshold or not.

The columns of Table 15 represent the six possible priority orders for attributes A, B, and C. The rows indicate 18 combinations of assumptions concerning whether each contrast in attributes is large enough to be decisive or not; i.e., $u(a'_F) - u(a'_G) \geq \Delta_A$, $v(b'_F) - v(b'_G) \geq \Delta_B$, and $t(c_G) - t(c_F) \geq \Delta_C$ are either true ("yes") or not ("no").

There are only five response patterns consistent with the 108 LS models generated by combining priority orders and threshold assumptions in Table 15: *FFFF*, *GFFF*, *GFGF*, *GGFF*, or *GGGG*. The pattern, *GFFF*, should not be confused as evidence of integration; this pattern merely indicates that if either attribute A or B favors F, the response is F. The pattern, *GGGF*, is not compatible with any of the LS models analyzed in Table 15; this pattern indicates that only when both A and B favor F, the response is F.

Because the labeling of the attributes (A, B, and C) is arbitrary, it should be clear that the analysis in Table 15 also shows that attributes B and C (with levels of A fixed) should show independence, as should attributes A and C (with levels of B fixed). That is, each pair of attributes should show integrative independence according to the family of LS models. In addition, we can exchange the roles of F and G in the test.

Not all of the factorial tests in Study 2 match these assumptions concerning the ten levels that define each test; therefore, in each test, all possible LS models are worked out by the same method as in Table 15, providing a proof for each of the tests employed. For example, the predictions for the test in Table 4 are shown in Table 16, which departs from the assumptions used in Table 15 (note that two attributes favor R in the first choice). Nevertheless, Table 16 shows that the design in Table 4 is a proper test of integration because none of the LS models predicts the pattern *RRRS*. The same method is used to analyze Table 5, which is a combined test of integration and interaction, and Tables 6 and 7, which are proper tests of integration. The study in Table 8 is a partial test.

Appendix B. Proofs of properties in Table 1

The TAX model violates priority dominance, integrative independence, interactive independence, and satisfies transitivity.

Proof: To show that a model can violate a property, it suffices to find one instance of violation. In this paper, predictions are calculated for TAX using prior parameters and the predictions of the model are listed in Tables 2 and 3, showing violation of priority dominance in all three tests; predictions in Tables 4, 6 and 7 show that TAX implies integration for each pair of attributes. The TAX model violates interactive independence (calculated utilities of the

Table 15
Response patterns implied by lexicographic semiorder models; "no" in the first column means $0 < u(a_F) - u(a_G) < \Delta_A$; "yes" indicates $u(a_F) - u(a_G) \geq \Delta_A$. "Yes" in the second and third columns means $u(a'_F) - u(a'_G) \geq \Delta_A$ and $v(b_F) - v(b_G) \geq \Delta_B$ respectively; "Yes" in the fifth column means $t(c_G) - t(c_F) \geq \Delta_C$. Only five patterns appear: *FFFF*, *GFFF*, *GFGF*, *GGFF*, and *GGGG*. The pattern *GGGF* is not compatible with any of these 108 lexicographic semiorders.

(a_F, a_G)	(a'_F, a'_G)	(b_F, b_G)	(b'_F, b'_G)	(c_G, c_F)	ABC	ACB	BAC	BCA	CAB	CBA
No	No	No	No	No	*GGGG*	*FFFF*	*GGGG*	*FFFF*	*FFFF*	*FFFF*
No	No	No	No	Yes	*GGGG*	*GGGG*	*GGGG*	*GGGG*	*GGGG*	*GGGG*
No	No	No	Yes	No	*GFGF*	*FFFF*	*GFGF*	*FFFF*	*FFFF*	*FFFF*
No	No	No	Yes	Yes	*GFGF*	*GGGG*	*GFGF*	*GFGF*	*GGGG*	*GGGG*
No	No	Yes	Yes	No	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*
No	No	Yes	Yes	Yes	*FFFF*	*GGGG*	*FFFF*	*FFFF*	*GGGG*	*GGGG*
No	Yes	No	No	No	*GGFF*	*FFFF*	*GGFF*	*FFFF*	*FFFF*	*FFFF*
No	Yes	No	No	Yes	*GGFF*	*GGFF*	*GGFF*	*GGGG*	*GGGG*	*GGGG*
No	Yes	No	Yes	No	*GFFF*	*FFFF*	*GFFF*	*FFFF*	*FFFF*	*FFFF*
No	Yes	No	Yes	Yes	*GFFF*	*GGFF*	*GFFF*	*GFGF*	*GGGG*	*GGGG*
No	Yes	Yes	Yes	No	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*
No	Yes	Yes	Yes	Yes	*FFFF*	*GGFF*	*FFFF*	*FFFF*	*GGGG*	*GGGG*
Yes	Yes	No	No	No	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*
Yes	Yes	No	No	Yes	*FFFF*	*FFFF*	*FFFF*	*GGGG*	*GGGG*	*GGGG*
Yes	Yes	No	Yes	No	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*
Yes	Yes	No	Yes	Yes	*FFFF*	*FFFF*	*FFFF*	*GFGF*	*GGGG*	*GGGG*
Yes	Yes	Yes	Yes	No	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*FFFF*
Yes	Yes	Yes	Yes	Yes	*FFFF*	*FFFF*	*FFFF*	*FFFF*	*GGGG*	*GGGG*

Table 16
Lexicographic semiorder (LS) model analysis of test in Table 4. In this test, the four choices are R = (\$100, 0.9; \$50) versus S = (\$51, 0.5; \$50), R' = (\$100, 0.1; \$50) versus S' = (\$51, 0.5; \$50), R'' = (\$100, 0.9; \$0) versus S'' = (\$51, 0.5; \$10), and R''' = (\$100, 0.1; \$0) versus S''' = (\$51, 0.5; \$10). The preference pattern $R \succ S$, $R' \succ S'$, $R'' \prec S''$, and $R''' \prec S'''$ is denoted as *RRSS*. Question marks (?) denote choices where the LS model is undecided. Intervals indicate the range of values of the assumed parameter values. For example, the interval (0, 0.4] indicates that $0 < \Delta_P \leq 0.4$, and $(0.4, \infty)$ indicates that $0.4 < \Delta_P$ (i.e., is assumed that a difference of 0.4 in probability is either decisive or not.) The pattern predicted by the TAX model with prior parameters is *RRRS*; none of these LS models makes the same prediction.

Intervals of threshold parameters			Lexicographic priority order					
Δ_L	Δ_P	Δ_H	LPH	LHP	PLH	PHL	HPL	HLP
(0, 10]	(0, 0.4]	(0, 49]	*RSSS*	*RRSS*	*RSRS*	*RSRS*	*RRRR*	*RRRR*
(0, 10]	(0, 0.4]	$(49, \infty)$	*RSSS*	*RSSS*	*RSRS*	*RSRS*	*RSRS*	*RSSS*
(0, 10]	(0.4, 1)	(0, 49]	*RRSS*	*RRSS*	*RRSS*	*RRRR*	*RRRR*	*RRRR*
(0, 10]	(0.4, 1)	$(49, \infty)$	*RRSS*	*RSSS*	*RRSS*	*??SS*	*??SS*	*RSSS*
$(10, \infty)$	(0, 0.4]	(0, 49]	*RSRS*	*RRRR*	*RSRS*	*RSRS*	*RRRR*	*RRRR*
$(10, \infty)$	(0, 0.4]	$(49, \infty)$	*RSRS*	*RSRS*	*RSRS*	*RSRS*	*RSRS*	*RSRS*
$(10, \infty)$	(0.4, 1)	(0, 49]	*RRRR*	*RRRR*	*RRRR*	*RRRR*	*RRRR*	*RRRR*
$(10, \infty)$	(0.4, 1)	$(49, \infty)$	*RRRR*	*RSRS*	*RRRR*	*??SS*	*??SS*	*RSRS*

gambles are shown in Table 9). Because the TAX model represents the utility of a gamble with a single number, as in Eq. (1), it implies transitivity, apart from random error, by the same proof given after Eq. (1).

EU model is a special case of TAX and it is also a special case of CPT (Birnbaum, 2008b). With $u(x) = x^{\beta}$, where $\beta = 0.2$, EU makes the same predictions as prior TAX for Tables 2, 4–8 and 11. It predicts the pattern *SSSRR* for Table 9, violating interactive independence. Because EU is a special case of CPT, it shows that CPT is also compatible with these results. However, Birnbaum (2008b) has summarized strong evidence against EU and CPT.

The one attribute model satisfies priority dominance.

Proof: According to this model, people examine only one attribute and make their decisions based entirely on that attribute alone; therefore, the one attribute used is dominant. For example, if people choose by the lowest consequence alone, then that attribute has priority, and no change in the probability or the highest consequence can overcome it because the person does not use those attributes at all.

The one attribute model satisfies integrative independence.

Proof: Because a person chooses by one attribute, changes in other attributes have no effect. If variables have no effect, they cannot combine to change a decision. In a test of integration (Expressions (11a)–(11d)), one attribute has fixed levels and contrasts on two other attributes are varied. If the attribute that is fixed is the one attribute attended to, then manipulations of the other two attributes will have no effect, so the person would show the patterns, *BBBB*, *AAAA* or *????*. If the attribute used is one of the two manipulated in a factorial design, then the person can show the patterns *BBBB*, *BABA*, *BBAA*, *AAAA* or *????*, but will not show the pattern *BBBA*.

The one attribute model shows no interaction.

Proof: In this test, probability is held constant in the two gambles and there are differences on both the lower and higher consequences. If probability is the one attribute used, the person will show no preference in all of the choices in the test. If higher consequence is the one examined, then the person will choose the risky gamble in all cases in Table 9. If the lower consequence is the one examined, then the person will choose the "safe" gamble in all choices. Therefore, this model implies only the response patterns, *?????*, *RRRRR*, or *SSSSS*, but it cannot predict patterns such as *SSRRR*.

The one attribute model satisfies transitivity.

Proof: Because this model uses only one attribute, its utility can be represented as a function of that attribute's value. By the same argument given in association with Eq. (1), the transitivity of numerical utility implies the transitivity of preference.

The additive contrasts model (Eq. (4)) can violate priority dominance.

Proof: It suffices to find a violation. Suppose $\phi_1(p_{1G}, p_{1F}) = 122(p_{1G} - p_{1F})$, where $F = (x_{1F},\ p_{1F}; x_{2F})$ and $G = (x_{1G},\ p_{1G}; x_{2G})$, and suppose $\theta_1(x_{1G}, x_{1F}) = x_{1G} - x_{1F}$, for the higher consequences and $\theta_2(x_{2G}, x_{2F}) = 2(x_{2G} - x_{2F})$ for the lower consequences, where all other terms are set to zero. This model makes the same predictions as TAX for the three tests of priority dominance in Table 2; that is, it predicts *SRSRS* for the first six choices in Table 2, violating priority dominance.

The additive contrast model violates integrative independence (it shows integration).

Proof: With the assumptions in the previous paragraph, this model predicts *RRRS*, *SSSR*, *SSSR*, and *SSSR* for the tests of integration in Tables 4 and 6–8, the same as the prior TAX model.

The additive contrast model satisfies interactive independence (Expression (12)).

Proof: Because the probabilities are equal, the probability contrast terms are zero. It follows that:

$$D(A, B) = \sum_{i=1}^{2} \phi_i(p_i, p_i) + \sum_{i=1}^{2} \theta_i(x_{iA}, x_{iB}) = \sum_{i=1}^{2} \theta_i(x_{iA}, x_{iB})$$
$$= \sum_{i=1}^{2} \phi_i(p'_i, p'_i) + \sum_{i=1}^{2} \theta_i(x_{iA}, x_{iB}) = D(A', B').$$

Therefore, $A \succ B \Leftrightarrow A' \succ B'$. For this reason, this model cannot account for the results in Table 5, which tests a combination of integration and interaction, nor can it account for the results in Tables 9 and 10.

The additive contrast model violates transitivity, as shown by the example given in association with Eq. (5).

The family of models that includes regret theory and majority rule (Expression (6)) violates priority dominance, integrative independence, interactive independence and transitivity. To prove violations, it suffices to find a violation of each. To apply this model, we need to specify the set of mutually exclusive and exhaustive events on which the contrasts are defined. Let the events be defined as follows: E_1 = both gambles G and F win their higher prizes, $E_2 = G$ wins its higher prize and F wins its lower prize, $E_3 = G$ wins its lower prize and F wins its higher prize, and E_4 = both gambles win their lower prizes. Let $\phi_i(E_i) = p_G p_F$, $p_G(1 - p_F)$, $(1 - p_G)p_F$, and $(1 - p_G)(1 - p_F)$, for these four events, respectively. Let

$$\theta_i(x_{iG}, x_{iF}) = (x_{iG}^{\beta} - x_{iF}^{\beta})^{\gamma}$$

where $\beta = 0.5$ and $\gamma = 1.2$. With these assumptions, Eq. (6) makes the same predictions as TAX for Table 2, violating priority dominance; it makes the same predictions as TAX for Tables 4, 6 and 7, which shows that it violates integrative independence. It makes the same predictions as TAX for Tables 5 and 9, showing that it violates interactive independence. Although the model can predict violations of transitivity, this special case makes the same predictions as TAX for Tables 11 and 13, satisfying transitivity in these studies. This model also makes the same predictions as TAX for Table 8. Thus, Eq. (6) (which includes Regret Theory) can be retained as a description of the present results.

To show that Eq. (6) can violate transitivity, consider the example of three branch gambles given after Eq. (5). For gambles with three equally likely consequences, let $\phi(E_1) = \phi(E_2) = \phi(E_3) = 1/3$. Let θ_i be defined as in Eq. (5). Eq. (6) then implies intransitive choices for A = (\$80; \$40; \$30), B = (\$70; \$60; \$20), and C = (\$90; \$50; \$10). See Birnbaum and Schmidt (2008) for failed predictions of Regret Theory.

Appendix C. True and error model

As shown in Eq. (14), reversals between repeated presentations of the same choice provide an estimate of the error rate, $P(RS) + P(SR) = 2e(1 - e)$. For example, if there are 32% reversals on repeated presentations of a choice, the error rate is estimated to be 0.2.

The fact that $P(SS) = p(1 - e)^2 + (1 - p)e^2$ provides a method for estimating p. Note that if we know the error rate, which can be estimated from preference reversals across replications (Eq. (14)), we can use the rate of repeated choices (as in Eq. (13)) to find p.

In practice, p and e are estimated simultaneously to reproduce the four frequencies for each replicated choice. There are three degrees of freedom in these four frequencies (they sum to the number of participants), so there remains one degree of freedom to test the model. Using this method, the error terms for Choices 16, 13, 5, and 9 of Table 8 are estimated to be 0.05, 0.09, 0.21, and 0.15 respectively.

We can assess if $p = 0$ from the special case of $P(SS) = p(1 - e)^2 + (1 - p)e^2$, when $p = 0$. Note that when $e = 0.2$ and $p = 0$, we expect to see $P(SS) = 0.04$. This relatively small value shows that when the error rates are 0.2 or below, repeated choices are likely to be "real" rather than due to error.

Appendix D. Supplementary analyses of integration

D.1. Supplementary analysis of Table 4

To compare the family of LS models against the TAX model with free parameters, we can examine the predictions of these models. For the LS models, the patterns predicted are *RRRR*, *RRSS*, *RSRS*, and *RSSS*. The TAX model with free parameters can accommodate these 4 patterns, and it can also predict the pattern *RRRS*, when $(\beta, \gamma, \delta) = (1, 0.7, 1)$, which are the "prior" parameters. TAX predicts the patterns *RSSS*, *RSRS*, *RRSS* and *RRRR* when $(\beta, \gamma, \delta) = (0.2; 2.5, 1), (1.4, 2.2, 1), (0.5, 0.1, 1)$ and $(1.8, 0.7, 1)$, respectively. [Each response pattern is compatible with many combinations of parameters; these combinations are presented merely to show that the model can accommodate these patterns].

Therefore, in the test of Table 4, the family of LS is restricted to only four of the five patterns that are compatible with integrative models like TAX, CPT, and EU. With four choices, there are sixteen possible response patterns, ***RRRR***, ***RRRS***, *RRSR*, ***RRSS***, *RSRR*, ***RSRS***, *RSSR*, ***RSSS***, *SRRR*, *SRRS*, *SRSR*, *SRSS*, *SSRR*, *SSRS*, *SSSR*, *SSSS*. The number of people who showed each of these 16 patterns is **27**, **99**, 5, **21**, 8, **44**, 4, **8**, 0, 5, 3, 1, 3, 2, 2, and 10, respectively. (Patterns and frequencies set in bold font show cases consistent with TAX).

The "true and error" model of this four-choice property has 16 equations, each with 16 terms that describe the probability of showing each observed pattern and having each true pattern (combination of errors). For example, the probability of showing the observed pattern, *RRRS* and having the true pattern of *RRRR* is as follows:

$$P(RRRS \cap RRRR) = p(RRRR)(1 - e_1)(1 - e_2)(1 - e_3)e_4.$$

That is, a person whose true pattern is *RRRR* can show the *RRRS* pattern by correctly reporting the first three preferences and making an error on the fourth choice. There are fifteen other ways to show this response pattern, corresponding to the other fifteen possible true patterns of preference. The sum of these sixteen terms is the predicted probability of showing the pattern *RRRS*. The TAX model is a special case of this true and error model in which only five terms have non-zero probabilities (*RRRS*, *RSSS*, *RSRS*, *RRSS* and *RRRR*) and the LS family allows a still smaller subset in which the probability of *RRRS* is fixed to zero.

Because the 16 observed frequencies sum to the number of participants, the data have 15 degrees of freedom. For the TAX model with free parameters, there are five possible response patterns, so there are five "true" probabilities of the predicted patterns to estimate (which sum to 1, so four degrees of freedom are used) and four error rates to estimate (one for each choice). That leaves $15 - 4 - 4 = 7$ degrees of freedom. The family of LS models is a special case of this model in which $p(RRRS)$ is fixed to zero. The difference in χ^2 provides a test if the observed frequency of *RRRS* is significant, which would refute the LS family.

When this model was fit to minimize the $\chi^2(7)$, it yielded estimates of $(\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4)$ = (0.09, 0.10, 0.11, and 0.19); $\hat{p}(RSSS) = 0.00$, $\hat{p}(RSRS) = 0.44$, $\hat{p}(RRSS) = 0.00$, $\hat{p}(RRRS) = 0.53$, $\hat{p}(RRRR) = 0.03$. When $p(RRRS)$ was fixed to zero (testing the family of LS models), χ^2 increased by $\chi^2(1) = 11.81$, which

Table 17
True and error analysis of replicated design ($n = 266$) in Table 8. Entries shown in bold are consistent with integrative models such as TAX; *SSSR* is predicted by prior TAX. Estimated error terms are $(\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4) = (0.05, 0.09, 0.21, 0.15)$, respectively.

Response pattern	Number who show each pattern			Estimated probability
	Replicate 1	Replicate 2	Both replicates	
RRRR	7	5	1	0.05
RRRS	1	1	0	0
RRSR	5	4	0	0
RRSS	4	1	0	0
RSRR	3	1	0	0
RSRS	1	2	0	0
RSSR	5	3	0	0
RSSS	2	1	0	0
SRRR	7	2	0	0.00
SRRS	2	7	0	0
SRSR	26	25	14	0.12
SRSS	9	5	1	0
SSRR	48	52	18	0.14
SSRS	6	13	1	0
SSSR	93	99	52	0.48
SSSS	47	45	26	0.21

is significant. In sum, the observation that 99 people showed the *RRRS* pattern, which is not compatible with any of the LS models, cannot be reconciled with the true and error model.

Because many of the cells in this example had low predicted frequencies, the statistical tests were also repeated, pooling cells whose predicted frequencies were less than 4. Although values of χ^2 observed and degrees of freedom are altered in these analyses, the conclusions were not changed here or in other tests of this type. In this analysis $(\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4) = (0.04, 0.24, 0.11, 0.20)$, $\hat{p}(RRRS) = 0.77$, $\hat{p}(RRSS) = 0.06$, $\hat{p}(RSRS) = 0.09$, and $\hat{p}(SSSS) = 0.08$. All other parameters were set to zero. There were 7 cells that had predicted frequencies less than 4, which were pooled, leaving 10 cells with 9 df; the $\chi^2(2) = 0.87$, showing that this integrative model combined with the true and error model can be retained. The predicted frequencies from this model are 24.9, **100.8**, 5.1, **20.7**, 10.4, **41.9**, 2.0, 8.2, 1.3, 5.1, 1.0, 4.0, 0.8, 3.2, 2.5, and **10.1**, respectively (the four cells with nonzero probabilities in the model are shown in bold font). These predicted frequencies can be used to calculate four predicted choice percentages for Table 4.

When $\hat{p}(RRRS)$ was fixed to zero in this pooled analysis, the fit was significantly worse, $\chi^2(1) = 43.2$. In sum, the data are not compatible with the family of LS models even when cells with small predicted frequencies are pooled, but they can be reconciled by models that allow integration, such as the TAX model.

D.2. Supplementary analysis of Table 5

In this case, probability contrasts are always zero. There are two possibilities for the contrast in the lowest consequences ($0 versus $20 either reaches threshold or not); there are three possibilities for the threshold parameters in the highest consequences: either both $26 versus $25 and $100 versus $25 exceed threshold, or $26 versus $25 does not reach threshold but $100 versus $25 does, or both contrasts fall short of threshold. There are thus $2 \times 3 = 6$ assumptions concerning thresholds and six priority orders, yielding 36 LS models. These 36 LS models can predict the patterns, *RRRR*, *SRSR*, *SSSS*, *?R?R*, and *????*. Resolving the undecided cases in favor of either *S* or *R* results in just three response patterns. TAX model with its prior parameters predicts *SSSR*, a pattern that is not compatible with any of the LS models, but which was exhibited by 115 people. When its parameters are free, the TAX model can also accommodate the patterns *SSSS*, *SSRR*, *SRSR*, and *RRRR*, when $(\beta, \gamma, \delta) = (0.5; 0.1, 1)$, (2; 2.5, 0), (2; 0.1, 0), and (14; 1.2, 0), respectively. The family of LS models again allows only a subset of three of the five patterns consistent with TAX, in this case, including only three: *SSSS*, *SRSR*, and *RRRR*. The observed frequencies of the 16 possible patterns from ***RRRR***, *RRRS*, *RRSR*, *RRSS*, *RSRR*, *RSRS*, *RSSR*, *RSSS*, *SRRR*, *SRRS*, ***SRSR***, *SRSS*, ***SSRR***, *SSRS*, ***SSSR***, ***SSSS***, are **6**, 3, 5, 3, 2, 0, 3, 2, 12, 4, **25**, 2, **39**, 5, **115**, and **16**, respectively. Fitting the true and error model to these frequencies yielded the following solution: $\hat{p}(RRRR) = 0.07$, $\hat{p}(SRSR) = 0.19$, $\hat{p}(SSRR) = 0.0$, $\hat{p}(SSSR) = 0.74$, and $\hat{p}(SSSS) = 0.0$, with all others set to zero. When $\hat{p}(SSSR)$ was fixed to 0, the fit was significantly worse, as one might expect from the fact that 115 people showed this pattern, $\chi^2(1) = 39.55$. Pooling small frequencies, a simpler TAX model was also compatible with the data, $\hat{p}(RRRR) = 0.06$, $\hat{p}(SRSR) = 0.14$, and $\hat{p}(SSSR) = 0.80$, $\chi^2(3) = 5.87$. The predicted frequencies for this model are **10.6**, 2.0, 5.3, 1.0, 2.1, 0.4, 5.2, 1.0, 10.8, 2.1, 26.1, 5.0, **40.6**, 7.8, **102.2**, 19.7. From these predicted frequencies, one can calculate predicted frequencies of the four choices in Table 5.

D.3. Supplementary analysis of Table 6

The TAX model with free parameters can also predict the patterns *SSSS*, *SSRR*, and *SRSR* [when $(\beta, \gamma, \delta) = (0.6; 0.7, 1)$, (2; 2.5, 0), and (2.5; 0.5, 0.5), respectively]. Observed frequencies of the 16 response patterns from *RRRR*, *RRRS*, to *SSSS* are 2, 7, 1, 3, 1, 2, 3, 2, **6**, 2, **15**, 5, **25**, 2, **104**, and **62**, respectively. Fitting the true and error model, the estimated true probabilities are $\hat{p}(SRRR) = 0.06$, $\hat{p}(SRSR) = 0.04$, $\hat{p}(SSRR) = 0.12$, $\hat{p}(SSSR) = 0.69$, and $\hat{p}(SSSS) = 0.09$, with all others fixed to zero. This model predicts frequencies of 0.9, 0.3, 1.3, 0.6, 1.7, 0.7, 9.3, 4.6, **10.5**, 4.0, **14.5**, 6.5, **20.0**, 7.7, **107.2**, and **52.2**. The assumption that $p(SSSR) = 0$ can be rejected, $\chi^2(1) = 28.51$, which shows that the family of LS cannot handle these data.

D.4. Supplementary analysis of Table 7

The TAX model with free parameters can also handle the patterns *SSSR*, *SSSS*, and *SRSR* [with $(\beta, \gamma, \delta) = (1, 0.7, 1)$, (0.3, 0.7, 1), and (1.5, 0.7, 1), respectively]. The observed frequencies of individual response patterns from *RRRR* to *SSSS* were 2, 1, 4, 4, 1, 0, 2, 1, 4, 0, **80**, 12, **11**, 3, **92**, and **25**. The true and error model yielded the following estimates: $\hat{p}(RRRR) = 0.02$, $\hat{p}(SRSR) = 0.37$, $\hat{p}(SSRR) = 0.06$, $\hat{p}(SSSR) = 0.45$, and $\hat{p}(SSSS) = 0.10$, with all others set to zero. Setting $\hat{p}(SSSR) = 0$, the deviations were significant, $\chi^2(1) = 8.39$. A simpler model with $\hat{p}(SRSR) = 0.42$, $\hat{p}(SSSR) = 0.49$, $\hat{p}(SSSS) = 0.09$ and all others set to zero also provided an acceptable fit; $\chi^2(3) = 3.84$. In this simpler model, the predicted frequencies are 0.4, 0.1, 3.5, 0.6, 0.5, 0.1, 4.0, 1.2, 9.1, 1.6, **77.7**, 13.4, **10.5**, 3.1, **89.6**, **26.7**.

Appendix E. True and error model analysis of Table 8

The true and error model can be used for the analysis of an experiment testing a property involving four choices with replications, as in Table 8. The observed frequencies of showing each response pattern on first replicate, second replicate and on both replicates are shown in Table 17. According to the true and error model, each person may have one of sixteen "true" preference patterns, as in Appendix D. The probability of a person showing a given pattern of data is then the sum of sixteen terms, each representing the probability of showing a given response pattern and having the pattern of errors and correct responses required to produce that response pattern. There are sixteen equations, each of which has sixteen terms, one for each of the sixteen possible observed patterns. To calculate the probability of showing this pattern on both replicates, the terms representing correct choices and errors are squared; that yields another 16 equations for the probabilities of showing each pattern on two replications.

The true and error model was fit to the 16 frequencies of repeating a response pattern on two replications and the 16 average frequencies of showing these response patterns on either the first or second replication but not both. These 32 frequencies are mutually exclusive and sum to the number of participants, which means there are 31 degrees of freedom in the data. The substantive models to be compared are both special cases of this "true and error" model, in which many of the possible response patterns are assumed to have true probabilities of zero.

The TAX model with free parameters can handle 6 response patterns: *SSSS*, *SSSR*, *SSRR*, *SRSR*, *SRRR*, and *RRRR*. The probabilities of the other 10 patterns are set to zero. From the data, 4 error terms are estimated, and 6 "true" probabilities of response patterns are estimated. These 6 true probabilities sum to 1, so they use 5 degrees of freedom, leaving $31-4-5=22$ degrees of freedom to test the model. Fitting Table 17, $\chi^2(22) = 25.0$, an acceptable fit. The estimated "true" probabilities are shown in the last column of Table 17. One of the patterns compatible with TAX, $\hat{p}(SRRR)$ was estimated to be 0, but the pattern, *SSSR*, had a large estimated probability, $\hat{p}(SSSR) = 0.48$. This latter probability indicates that almost half of the sample has the response pattern that is predicted by TAX with its prior parameters and which is incompatible with the LS models except for PHL with $0.3 < \Delta_P \leq 0.5$.

When the probability of this pattern, *SSSR*, is set to zero, the best-fit solution yields $\chi^2(21) = 533.1$, an increase of $\chi^2(1) = 508.1$. In sum, whereas these data are compatible with the family of integrative models, they rule out all of the LS models except PHL2. However, because PHL2 is refuted by other tests in Studies 1–3, that model cannot be retained as descriptive, despite its compatibility with the data of Table 8.

Appendix F. True and error analysis of transitivity

The "true and error" model allows a person who is truly intransitive to show a transitive pattern and vice versa. For example, a person might correctly report preference for the first gamble (F) on Choices 6 and 18 and make an "error" on trial 14 by choosing the second gamble, S. The probability that this *FFS* response pattern would be observed, given that the person's true pattern is *FFF* is given as follows:

$$P(\text{observed} = FFS|\text{true} = FFF) = (1-e_1)(1-e_2)e_3 \quad (16)$$

where e_1, e_2, and e_3 are the error rates for Choices 6, 18, and 14, respectively. Assuming Choices 15, 12, and 8 are replicates of 6, 18, and 14, respectively, we can estimate error terms. The theoretical probability of each response pattern is the sum of eight terms, representing the eight mutually exclusive and exhaustive conjunctions of observed and "true" patterns.

The data are partitioned into the frequencies of showing each of 8 patterns on both replicates and the average frequency of showing each of these response patterns on one replicate or the other but not both. The model is fit to these 16 frequencies to minimize the χ^2 between predicted and obtained frequencies. The true and error model is neutral with respect to the issue of transitivity, since it allows each subject to have a different true pattern, which need not be transitive.

When this model was fit to the data, the best-fit solution yielded the following estimates for the "true" probabilities of the 8 possible response patterns: 0, 0.87, 0.03, 0, 0.03, 0, 0.07, 0 for *FFF*, *FFS*, *FSF*, *FSS*, *SFF*, *SFS*, *SSF*, and *SSS*, respectively, where F = choice of the First gamble and S = choice of the Second gamble in Table 11. Thus, this model estimates that 87% of the participants had *FFS* as their "true" preference pattern, and that no one had either the intransitive pattern predicted by the priority heuristic (*FFF*) or the opposite intransitive pattern (*SSS*) as his or her "true" pattern. The estimated error rates were 0.11, 0.09, and 0.08, for the first, second, and third choices in the table, respectively.

References

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropoliannae*, *5*, 175–192. (Translation by L.S. appears as Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. Econometrica, 22, 23–36).

Birnbaum, M. H. (1973). The Devil rides again: correlation as an index of fit. *Psychological Bulletin*, *79*, 239–242.

Birnbaum, M. H. (1974a). The nonadditivity of personality impressions. *Journal of Experimental Psychology*, *102*, 543–561.

Birnbaum, M. H. (1974b). Reply to the Devil's advocates: don't confound model testing and measurement. *Psychological Bulletin*, *81*, 854–859.

Birnbaum, M. H. (1997). Violations of monotonicity in judgment and decision making. In A. A. J. Marley (Ed.)., *Choice, decision, and measurement: Essays in honor of R. Duncan Luce* (pp. 73–100). Mahwah, NJ: Erlbaum.

Birnbaum, M. H. (1999). Testing critical properties of decision making on the Internet. *Psychological Science*, *10*, 399–407.

Birnbaum, M. H. (2004a). Causes of Allais common consequence paradoxes: an experimental dissection. *Journal of Mathematical Psychology*, *48*(2), 87–106.

Birnbaum, M. H. (2004b). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes*, *95*, 40–65.

Birnbaum, M. H. (2005a). A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *Journal of Risk and Uncertainty*, *31*, 263–287.

Birnbaum, M. H. (2005b). Three new tests of independence that differentiate models of risky decision making. *Management Science*, *51*, 1346–1358.

Birnbaum, M. H. (2008a). Evaluation of the priority heuristic as a descriptive model of risky decision making: comment on Brandstatter et al (2006). *Psychological Review*, *115*, 253–262.

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

Birnbaum, M. H. (2008c). New tests of cumulative prospect theory and the priority heuristic: probability-outcome tradeoff with branch splitting. *Judgment and Decision Making*, *3*, 304–316.

Birnbaum, M. H., & Bahra, J. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science*, *53*, 1016–1028.

Birnbaum, M. H., & Chavez, A. (1997). Tests of theories of decision making: violations of branch independence and distribution independence. *Organizational Behavior and Human Decision Processes*, *71*(2), 161–194.

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semiorder. *Organizational Behavior and Human Decision Processes*, *104*, 97–112.

Birnbaum, M. H., & Jou, J. W. (1990). A theory of comparative response times and "difference" judgments. *Cognitive Psychology*, *22*, 184–210.

Birnbaum, M. H., & LaCroix, A. R. (2008). Dimension integration: testing models without trade-offs. *Organizational Behavior and Human Decision Processes*, *105*, 122–133.

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes*, *67*, 91–110.

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

Birnbaum, M. H., & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty*, *37*, 77–91.

Blavatskyy, P. R. (2006). Axiomatization of a preference for most probable winner. *Theory and Decision*, *60*, 17–33.
Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: choices without tradeoffs. *Psychological Review*, *113*, 409–432.
Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008). Risky Choice with Heuristics: reply to Birnbaum (2008), Johnson, Schulte-Mecklenbeck, & Willemsen (2008) and Rieger & Wang (2008). *Psychological Review*, *115*, 281–290.
Busemeyer, J. R., & Townsend, J. T. (1993). Decision Field Theory: a dynamic cognition approach to decision making. *Psychological Review*, *100*, 432–459.
Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380–417.
Glöckner, A., & Betsch, T. (2008). Do people make decisions under risk based on ignorance? An empirical test of the priority heuristic against cumulative prospect theory. *Organizational Behavior and Human Decision Processes*, *107*, 75–95.
González-Vallejo, C. (2002). Making trade-offs: a probabilistic and context-sensitive model of choice behavior. *Psychological Review*, *109*, 137–155.
Iverson, G., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.
Iverson, G.J., & Myung, J.I. (2010). Intransitivity of preference: Revisited. Working paper. Available from Geoffrey J. Iverson, Institute for Mathematical Behavioral Sciences, UCI, Irvine, CA 92697.
Johnson, J. G., & Busemeyer, J. R. (2005). A dynamic, computational model of preference reversal phenomena. *Psychological Review*, *112*, 841–861.
Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*, 263–291.
Leland, J. W. (1994). Generalized similarity judgments: an alternative explanation for choice anomalies. *Journal of Risk and Uncertainty*, *9*, 151–172.
Leland, J. W. (1998). Similarity judgments in choice under uncertainty: a re-interpretation of the predictions of regret theory. *Management Science*, *44*, 659–672.
Lopes, L. L., & Oden, G. C. (1999). The role of aspiration level in risky choice: a comparison of cumulative prospect theory and SP/A theory. *Journal of Mathematical Psychology*, *43*, 286–313.
Loomes, G., Starmer, C., & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica*, *59*, 425–440.
Loomes, G., & Sugden, R. (1982). Regret theory: an alternative theory of rational choice under uncertainty. *Economic Journal*, *92*, 805–824.
Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, *24*, 178–191.
Luce, R. D. (2000). *Utility of gains and losses: Measurement-theoretical and experimental approaches*. Mahwah, NJ: Lawrence Erlbaum Associates.
Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent linear utility models for finite first order gambles. *Journal of Risk and Uncertainty*, *4*, 29–59.
Luce, R. D., & Fishburn, P. C. (1995). A note on deriving rank-dependent utility using additive joint receipts. *Journal of Risk and Uncertainty*, *11*, 5–16.
Marley, A. A. J., & Luce, R. D. (2001). Rank-weighted utilities and qualitative convolution. *Journal of Risk and Uncertainty*, *23*(2), 135–163.
Marley, A. A. J., & Luce, R. D. (2005). Independence properties vis-à-vis several utility representations. *Theory and Decision*, *58*, 77–143.
Quiggin, J. (1993). *Generalized expected utility theory: The rank-dependent model*. Boston: Kluwer.
Regenwetter, M., Dana, J., & Davis-Stober, C. (2010). Transitivity of Preferences. Psychological Review, in press.
Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: a review of research on preferential choice. *Journal of Economic Literature*, *35*, 631–661.
Rubinstein, A. (1988). Similarity and decision-making under risk (Is there a utility theory resolution to the Allais paradox?). *Journal of Economic Theory*, *46*, 145–153.
Sopher, B., & Gigliotti, G. (1993). Intransitive cycles: rational Choice or random error? An answer based on estimation of error rates with experimental data. *Theory and Decision*, *35*, 311–336.
Starmer, C. (2000). Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*, 332–382.
Stevenson, M. K., Busemeyer, J. R., & Naylor, J. C. (1991). Judgment and decision-making theory. In M. Dunnette, & L. M. Hough (Eds.), *New handbook of industrial-organizational psychology* (pp. 283–374). Palo Alto, CA: Consulting Psychologist Press.
Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.
Weber, B. J. (2007). The effects of losses and event splitting on the Allais paradox. *Judgment and Decision Making*, *2*, 115–125.
Wu, G., Zhang, J., & Gonzalez, R. (2004). Decision under risk. In D. Koehler, & N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making* (pp. 399–423). Oxford: Oxford University Press.
Zhang, J., Hsee, C. K., & Xiao, Z. (2006). The majority rule in individual decision making. *Organizational Behavior and Human Decision Processes*, *99*, 102–111.