Psychological Review
2008, Vol. 115, No. 1, 253–262

0033-295X/08/$12.00 DOI: 10.1037/0033-295X.115.1.253

Evaluation of the Priority Heuristic as a Descriptive Model of Risky Decision Making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006)

Michael H. Birnbaum
California State University, Fullerton

E. Brandstätter, G. Gigerenzer, and R. Hertwig (2006) contended that their priority heuristic, a type of lexicographic semiorder model, is more accurate than cumulative prospect theory (CPT) or transfer of attention exchange (TAX) models in describing risky decisions. However, there are 4 problems with their argument. First, their heuristic is not descriptive of certain data that they did not review. Second, their analysis relied on a global index of fit, percentage of correct predictions of the modal choice. Such analyses can lead to wrong conclusions when parameters are not properly estimated from the data. When parameters are estimated from the data, CPT and TAX fit the D. Kahneman and A. Tversky (1979) data perfectly. Reanalysis shows that TAX and CPT do as well as the priority heuristic for 2 of the data sets reviewed and outperform the priority heuristic for the other 3. Third, when 2 of these sets of data are reexamined, the priority heuristic is seen to make systematic violations. Fourth, new critical implications have been devised for testing the family of lexicographic semiorders including the priority heuristic; new results with these critical tests show systematic evidence against lexicographic semiorder models.

Keywords: choice, descriptive models, heuristics, risky decision making, utility

Brandstätter, Gigerenzer and Hertwig (2006) presented the priority heuristic as a descriptive model of risky decision making. This model is a variant of a lexicographic semiorder, such as that applied by Tversky (1969) and others. They consider such models "fast and frugal" because choices can be made, in some cases, without examining all of the information. They reanalyzed some published data and concluded that their model fits those data better than parametric models such as Tversky and Kahneman's (1992) cumulative prospect theory (CPT) and Birnbaum's (1997, 1999b, 2004a; Birnbaum & Chavez, 1997; Birnbaum & Navarrete, 1998) transfer of attention exchange (TAX) model.

What can one make of this supposedly "good" fit of the priority heuristic to previous data? My contention is that Brandstatter et al. (2006) were somewhat selective in the data they described and that their procedures for data analysis are questionable. Examination of other data and reanalysis of the data they reviewed cast doubt on the descriptive accuracy of their model. Their approach of stating that the model is not applicable in cases in which it fails is also problematic. As it turns out, the conditions they list to exclude tests can be seen as reducing the domain of their theory to a very small region, and even within that restricted region, the data depart significantly from predictions of their model. New tests have been devised to test implications of the family of lexicographic semiorders, and these new data violate the model's implications.

Selective Data Review

Brandstätter et al. (2006) conceded that their priority heuristic does not account for the dissection of the Allais paradox (Birnbaum, 2004a), but they did not describe a number of other choices in which the priority heuristic also fails to predict the results. For the 13 cases in Tables 2, 3, and 4 of Birnbaum (2004a), in which the choice percentage was significantly different from 50%, the priority heuristic correctly predicted the modal choice in only three cases. In four cases, it made erroneous predictions, and in six cases it made no prediction. The TAX model (with parameters estimated from previous data) correctly predicted the modal choice in all 13 cases.

Although Brandstätter et al. (2006) cited Birnbaum and Navarrete (1998), they did not mention that their model correctly predicted the modal choice in only 49 of 112 choices reported in Birnbaum and Navarrete, whereas the TAX model with prior parameters correctly predicted 89 of those choices. For example, most participants in Birnbaum and Navarrete should have chosen the "risky" gamble, R = ($98, 0.1; $10, 0.1; $3, 0.8) in preference to ($>$) the "safe" gamble, S = ($52, 0.1; $48, 0.1; $3, 0.8), according to the priority heuristic, because lowest consequences and probabilities are the same, so people should have chosen R, with the higher best consequence. And most people should have chosen S' = ($107, 0.8; $52, 0.1; $48, 0.1) $>$ R' = ($107, 0.8; $98, 0.1; $10, 0.1) because S' has the better lowest consequence. Instead, significantly more participants in that study had the op-

This research was supported by National Science Foundation Grants BCS-0129453 and SES-0202448. Thanks are due to Eduard Brandstätter for providing his data files for reanalysis and for fruitful discussions of these issues and to Jerome Busemeyer and Anthony J. Bishara for their comments.

Correspondence concerning this article should be addressed to Michael H. Birnbaum, Department of Psychology, California State University, Fullerton, H-830M, PO Box 6846, Fullerton, CA 92834-6846. E-mail: mbirnbaum@fullerton.edu

posite combination of preferences (30) than showed the pattern consistent with this heuristic (8), $z = 3.57$.

Similarly, only 31% of 100 participants in Birnbaum and Navarrete's (1998) study chose R_1 = (\$97, 0.1; \$11, 0.1; \$2, 0.8) over S_1 = (\$56, 0.1; \$52, 0.1; \$2, 0.8), even though the highest consequence of R_1 is more than \$10 greater than that of S_1. Furthermore, 73% chose R_2' = (\$110, 0.5; \$96, 0.25; \$12, 0.25) over S_2' = (\$110, 0.5; \$34, 0.25; \$30, 0.25), despite the fact that the lowest consequence of S_2' is more than \$10 greater than that of R_2'. This pattern of significant, incorrect prediction occurred in many other cases in Birnbaum and Navarrete (1998), as well as in earlier publications (Birnbaum & Chavez, 1997; Birnbaum & McIntosh, 1996).

Stochastic Dominance

The priority heuristic fails to predict satisfactions of stochastic dominance in cases where most people satisfy this property. For example, Birnbaum (1999b) reported that 92% chose H = (\$100, 0.2; \$96, 0.3; \$4, 0.5) over I = (\$100, 0.2; \$12, 0.3; \$4, 0.5), even though the priority heuristic cannot resolve this choice. Similarly, the priority heuristic cannot decide that D = (\$100, 0.90; \$30, 0.10) $\succ$ C = (\$100, 0.82; \$22, 0.18). The lowest consequences differ by only \$8, so people would next examine probabilities, which differ by only 0.08, so they would look at the highest consequences, which are the same, and so be unable to decide.

The priority heuristic also fails to predict violations of stochastic dominance in cases in which most undergraduates violate it (Birnbaum, 1999b, 2005a; Birnbaum & Navarrete, 1998). In a new variation, 71% of 408 undergraduates tested for this comment chose F = (\$89, 0.7; \$88, 0.1; \$11, 0.2) over G = (\$90, 0.8; \$13, 0.1; \$12, 0.1), despite the fact that G stochastically dominates F. In this example, G has a higher lowest outcome, higher best outcome, lower probability to get the worst consequence, higher probability to get the best consequence, and a higher expected value (EV). G should be chosen by the priority heuristic because of its lower probability to receive the lowest outcome. The only way in which F is better than G is in the consequences on the middle branch, which the priority heuristic ignores. Birnbaum (2005a) found that participants indeed respond significantly to manipulation of the middle consequence. To avoid such wrong predictions, Brandstätter et al. (2006) stated that their heuristic does not apply to choices with a stochastic dominance relation.

Excluding cases of dominance creates four problems for priority heuristic as a descriptive theory. (a) It creates a theoretical problem. How do people perceive dominance? To decide not to use the priority heuristic, people must be able to perceive dominance, and to do this, they must examine all of the information in both gambles, so adding this preliminary decision stage contradicts the goal of being "fast and frugal." (b) If people can detect dominance, then why not obey it? (c) It is all too easy to say that a theory is perfect, except where it does not apply—and then add to the list of exceptions. (d) This restriction of the domain of the theory cuts down the applicability of the priority heuristic considerably.

Consider two-branch gambles of the form $A = (x, p; y, 1 - p)$, where \$100 $\geq x, y \geq 0$. Suppose we sample by the following procedure: Choose x and y by randomly drawing from a uniform distribution from \$0 to \$100. Choose p by randomly sampling from a uniform distribution on the interval from 0 to 1. Now choose pairs of such "random" gambles independently. In this domain, one half of all potential choices are excluded by the boundary condition excluding dominance. Half the space seems a large region to exclude. Now consider choices between cash, c, and such binary gambles, where c is uniformly distributed on the same interval: In this situation, stochastic dominance excludes two thirds of the potential choices!

New Paradoxes

Brandstätter et al. (2006) conceded that their heuristic does not predict Birnbaum's "new paradoxes," which involve two- and three-branch gambles. Those paradoxes were designed to refute CPT without having to assume or estimate any parameters. Like CPT, the priority heuristic does not correctly describe violations of upper and lower cumulative independence (Birnbaum, 1999b, 2004b; Birnbaum & Navarrete, 1998). Of the 12 choices analyzed by Birnbaum (1999b, Table 1) to compare CPT and TAX, the priority heuristic is correct in only 4 of 12 choices: It makes no prediction in four cases, and it makes wrong predictions in four other cases. In 17 choices analyzed by Birnbaum (1999a), the priority heuristic correctly predicts the modal choice in only five cases, it is wrong in eight cases, and it makes no prediction in four cases. Birnbaum (1999a) showed that one can reproduce those 17 choices with the TAX model using the approximation, $u(x) = x$; but this does not mean that only this utility function works.

The priority heuristic cannot account for systematic violations of distribution independence (Birnbaum, 2005b; Birnbaum & Chavez, 1997). In sum, there is a considerable body of previously published evidence showing that this heuristic is not an accurate descriptive model, but these data were not included by Brandstätter et al. (2006) in their contests of fit. Including these data leads one to the conclusion that the priority heuristic is not an accurate descriptive model. If we theorize that people use a preliminary decision process to decide not to use the priority heuristic for these cases, we need to devise a mechanism that can detect these choices for exclusion.

Data Analysis

The analysis of the data of Tversky and Kahneman (1992) in Brandstätter et al. (2006) was not diagnostic. Tversky and Kahneman studied choices between sure cash, c, and binary gambles of the form, $(x, p; y)$, where $x > y$. The data were published in the form of certainty equivalents (CEs), which are values of c such that people preferred the cash or the gamble 50% of the time. Brandstätter et al. (2006) analyzed those data by asking how often each model correctly predicted the relationship between the CE and the EV. By this criterion, CPT, TAX, and the priority heuristic all seem to do well, but this criterion is not diagnostic of differences among the models.

If, instead of comparing CE with EV, we use 60% of the EV, we can better distinguish the models. According to CPT and TAX, with their prior parameters, people should choose the gamble over 0.6 EV when $x > y \geq 0$ for all $p > 0.2$. According to the priority heuristic, however, people should choose the sure cash when it exceeds the rounded value of 10% of the largest consequence—that is, when $p \geq 0.2$ and $c \geq 0.6$ EV. (The reason to use 60% of EV is to allow the models to make different predictions and also to

ensure that choices are inside the region estimated by Brandstätter et al., 2006, where the priority heuristic is supposed to apply; i.e., $0.5 < EV/c < 2$.)

In Table 3 of Tversky and Kahneman (1992), there are 20 "choices" between c and (x, p; y) that fit these diagnostic criteria (i.e., $p \geq 0.2$ and $c \geq 0.6$ EV). For example, according to the priority heuristic, people should prefer \$30 over (\$100, 0.5; \$0) because \$30 exceeds 10% of \$100. However, CPT and TAX, with their prior parameters, imply that people should prefer the gamble. Tversky and Kahneman found that most people chose the gamble over \$30, resulting in a CE of \$36 in this case. Of these 20 diagnostic choices, both TAX and CPT are correct in 100% of the cases, and the priority heuristic is correct in 0% of the cases. Similar results are observed with strictly nonpositive consequences. These conclusions are quite different from those reached by Brandstätter et al. (2006), who argued that the priority heuristic accurately predicts these data; instead, reanalysis shows that we can reject the priority heuristic as a description of Tversky and Kahneman's data in favor of TAX or CPT.

Rieger and Wang (in press) used an index of fit to show that CPT fits the Tversky and Kahneman (1992) data better than does the priority heuristic. Because TAX makes predictions that are virtually the same as those of CPT for that experiment, TAX will also fit those data better than the priority heuristic by that same index of fit.

Parameter Estimation

The contests of fit in Brandstätter et al. (2006) did not allow parameter estimation to models that use parameters. For example, Figure 1 in Brandstätter et al. displays a comparison of the accuracy of models to 14 choices from Kahneman and Tversky (1979). According to the figure, the priority heuristic predicts the modal decision in 100% of the 14 choices, whereas TAX and CPT are supposedly correct in only 70% of the 14 cases. However, TAX predicts those choices perfectly if it is allowed a nonlinear utility function, $u(x) = x^{\beta}$, where $\hat{\beta} = 0.7$, and the other parameters are the same as those used by Brandstätter et al. CPT can also fit those data perfectly, if it is allowed to estimate its parameters from those data.

Parametric models do not assume that every person has the same parameters nor do they assume that every experiment will induce the same parameters. Indeed, there are significant differences in choice behavior between men and women, between the highly educated and the less educated, and between individuals in the same homogeneous group (Birnbaum, 1999b, 2006). Brandstätter et al. (2006) agreed that their model cannot account for such results unless it employs free parameters.

In the TAX model, the approximation, $u(x) = x$, provided a reasonable approximation to group data of American undergraduates in experiments involving small variation of positive cash prizes less than \$150. Brandstätter et al. (2006) assumed that this approximation should also be appropriate for choices involving prizes ranging up to 2 months' salary for Israelis (Kahneman & Tversky, 1979). Although the linear approximation has been useful for simplifying the exposition of the TAX model (e.g., Birnbaum, 1999a), this approximation is not part of the model (Birnbaum, 1999a, p. 48), and it is not optimal.

When Birnbaum and Navarrete (1998) fit the TAX model with $u(x) = x^{\beta}$, with prizes less than \$150, the best fit to averaged data led to $\hat{\beta} = 0.73$. When TAX was fit to individuals, the median best-fit value was 0.41. Birnbaum and Chavez (1997) reported the median best fit value was $\beta = 0.61$. Therefore, the approximation, $u(x) = x$, is not an optimal fit even for the majority of undergraduates tested. Brandstätter et al. (2006) cited these two articles, but they did not use best-fit parameters from those articles; instead, they assumed that $\beta = 1$ in the TAX model. Because the Kahneman and Tversky (1979) data can be fit perfectly by TAX and CPT, as well as by the priority heuristic, those data are simply not diagnostic for comparing these theories.

The Lopes and Oden (1999) data with five branch gambles can be fit by TAX with $\hat{\beta} = 0.41$ and with the other parameters fixed to the values used by Brandstätter et al. (2006). In this case, TAX correctly predicts 87% of the choices, the same as for CPT and the priority heuristic. So these data also provide no reason to prefer one of these models over the others.

Erev, Roth, Slonim, and Barron (2002) studied choices between binary gambles of the form (x, p; 0). Figure 5 of Brandstätter et al. (2006) shows that the priority heuristic is more accurate than TAX when $\beta = 1$. The priority heuristic, with its prior parameters, had 15 errors out of 100 choices (85% correct). The solid curve in Figure 1 shows the percentage of wrong predictions of the TAX model as a function of β, where the other parameters are the same as in Brandstätter et al. Figure 1 shows that TAX fits better than the priority heuristic when $0.06 \leq \beta \leq 0.68$. With $\hat{\beta} = 0.31$, TAX

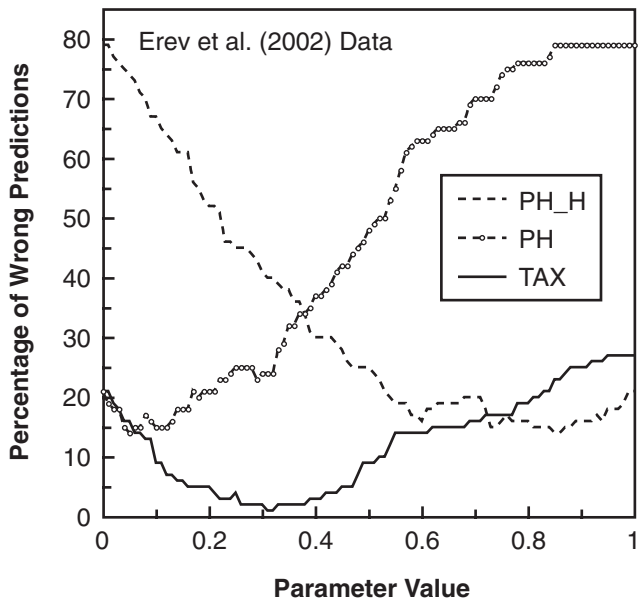

Figure 1. Percentage of wrong predictions in the Erev et al. (2002) data, plotted as a function of the parameter, with a separate curve for each model. For the transfer of attention exchange model (TAX), the parameter is β, the exponent of a power function for utility, $u(x) = x^{\beta}$. For priority heuristic (PH), the parameter is the threshold for probability. For the priority model in which the highest consequence has priority (PH_H), the parameter is the threshold factor for the highest consequence (i.e., the threshold is this parameter multiplied by the maximum of the higher consequences in the choice).

is correct in all but 1 of the 100 choices. Brandstätter et al. noted that CPT also fits these data with only one error when its parameters are estimated from the data.

For comparison, the priority heuristic with two free parameters (priority order and threshold parameter) was fit to the same data. Assuming probability has priority over the highest consequence (see circles connected by broken lines in Figure 1), the best-fit value of the probability threshold is 0.05, in which case the model is wrong in 14% of the choices. Assuming that highest consequence has priority (see dashed curve in Figure 1), this version of the heuristic achieves a best-fit value of 14% errors, with a threshold = 0.86 max(x, x'). If one should conclude anything from these data, it is that the best-fitting TAX and CPT models are more accurate than the best-fitting priority heuristic model.

Mellers, Chang, Birnbaum, and Ordóñez (1992) also studied choices between binary gambles of the form (x, p; 0) constructed from a 6 × 6 factorial design of x and p. The choice proportions can be approximated with a strongly transitive utility model with logistic function

$$P(A, B) = 1/\{1 + \exp(\alpha[u(A) - u(B)])\},$$

where $u(A)$ and $u(B)$ are the utilities of gambles and α is the logistic spread parameter. The least squares solution predicted the modal choice in 421 out of 450 choices (6.4% errors). When the above model was further constrained so that the gamble utilities—$u(A)$ and $u(B)$—satisfied the TAX model with one free parameter (β), the model correctly reproduced 407 out of 450 choices (9.5% errors), with $\hat{\beta} = 0.48$, and with the other parameters fixed to their prior values. Figure 2 shows the percentage of wrong predictions for the same three models, as in Figure 1.

The priority heuristic (with its parameter of 0.1) correctly predicted only 327 of the same 450 choices (27% errors), and its errors were highly systematic. For example, the priority heuristic implies that most people should choose $S = (x, 0.29; 0)$ over $R = (x', 0.17; 0)$ for any values of x and $x' > 0$; this prediction is wrong for all 30 of these choices in Mellers et al. (1992). In addition, the priority heuristic implies violations of weak stochastic transitivity that did not materialize in the data. For example, 82% chose A = (\$17.5, 0.05; \$0) over B = (\$5.4, 0.09; \$0) and 55% chose B over C = (\$3, 0.17; \$0), but 76% chose A over C, even though the priority heuristic predicts that the majority of participants should have chosen C over A. The EVs of A, B, and C are 0.85, 0.49, and 0.51, respectively, well inside the region in which the heuristic is supposed to work. The best-fit priority heuristic had its probability threshold = 0.23, in which case it made 19% errors.

Brandstätter et al. (2006) noted that their model was not accurate when EVs differed by a ratio greater than two. Excluding these cases, the priority heuristic makes predictions for only 266 of the 450 choices in Mellers et al (1992). Among these 266 cases, the priority heuristic had 21% wrong predictions, and it was still wrong in predicting all 18 of the 18 remaining choices between $S = (x, 0.29; 0)$ and $R = (x', 0.17; 0)$. When the best-fit two-parameter priority heuristic was fit inside the EV-restricted region, the best-fit probability threshold was 0.18, with 17% errors. These data therefore fail to show any advantage for the priority heuristic over TAX because the best-fit TAX model gives a better fit to all of the data (9.5% errors for 450 predictions) than does the best-fitting priority heuristic in the selected data (17% errors for 266 predictions). Similar results were observed in an analysis of the strictly nonpositive gambles of Mellers et al.

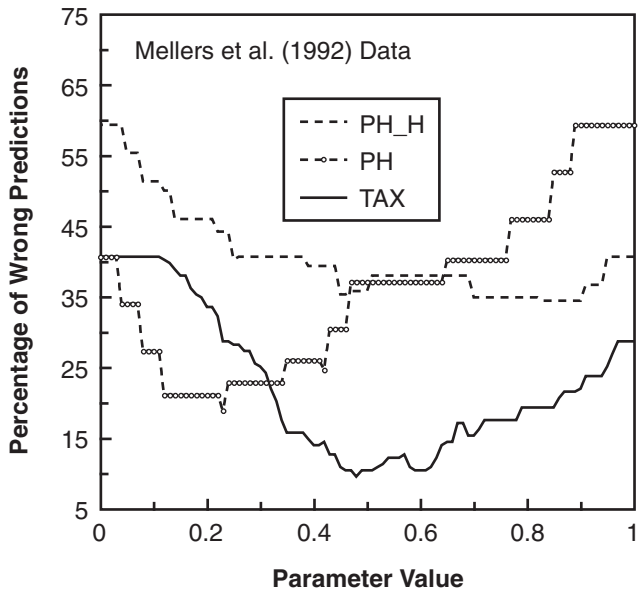

Figure 2. Fit to Mellers et al. (1992) data, plotted as a function of the parameter, with a separate curve for each model. For the transfer of attention exchange model (TAX), the parameter is β, the exponent of a power function for utility, $u(x) = x^{\beta}$. For priority heuristic (PH), the parameter is the threshold for probability. For the priority model in which the highest consequence has priority (PH_H), the parameter is the threshold factor for the highest consequence (i.e., the threshold is this parameter multiplied by the maximum of the higher consequences in the choice). The TAX achieves a better fit than the PH and the PH_H.

In sum, reanalysis shows that TAX and CPT perform as well or better than the priority heuristic for all five sets of data analyzed by Brandstätter et al. (2006) if TAX is allowed to use $\beta < 1$ instead of $\beta = 1$ and CPT is allowed to estimate its parameters from the data. In two of the data sets, the fit is the same for TAX, CPT, and the priority heuristic; in two cases, the priority heuristic makes systematic errors and can be rejected in favor of TAX or CPT. In the fifth case (Erev et al., 2002), TAX and CPT fit better than the best-fitting priority heuristic.

To show that a model is compatible with a set of data, it suffices to find a set of parameters that allow that model to reproduce those data. However, to disprove a model, one must show that there are no parameters in that model that allow it to fit the data. The standard Allais paradoxes do just that: When one attempts to define a utility scale from the paradoxical choices, expected utility theory leads to a contradiction. Brandstätter et al. (2006) argued, instead, that one should be able to refute a model by showing that parameters that worked in one experiment are not optimal in another experiment with new participants in a new context.

Brandstätter et al. (2006) conceded that all models have parameters, but they said that their parameter, 1/10, for the probability threshold and aspiration level was "derived" from our cultural base-10 number system. For the data of Mellers et al. (1992), the best-fit parameter for the heuristic was about 1/5, which could be "derived" by counting fingers and thumb of the right hand. For data of Erev et al. (2002), the best-fit value is close to 1/20, which

could be "derived" by counting on all fingers and toes. Similarly, the value 1/2 could be "derived" from the number of eyes in a head, 1/3 from the Holy Trinity, 1/4 from the directions on a compass, 1/7 from the number of days in a week, 1/8 by counting fingers but not thumbs, 1/12 from the number of hours on the face of a clock, 1/24 from the number of hours in a day, and so on. Unless one plans to randomly assign people to different cultures in which these cultural constants are manipulated (e.g., changing the cultural number system), postulating which of these cultural values is "the" cause of a parameter value strikes me as idle speculation.

Global Index of Fit

A global index of fit is simply not a good tool for comparing rival models. "High" indices of fit often coexist with serious, systematic discrepancies for a model. It has been shown that correlation coefficients between theory and data can even be higher for models that give worse descriptions of the data when measurement assumptions are confounded with model fitting by the use of a priori parameters (Birnbaum, 1973, 1974b). The same problem for correlation coefficients applies when one uses percentage correct as the global index of fit and confounds measurement and model testing by not estimating parameters within the model being tested. For example, consider the perfectly multiplicative data from Example 1 of Birnbaum (1973, 1974b). Suppose one categorizes a prediction as "correct" when a model correctly predicts whether the dependent variable is greater or less than 28.5. It was found that the additive model was correct in 96% of the predictions, and the multiplicative model had only 75% correct. This example illustrates that the methods used by Brandstätter et al. (2006) lead to wrong conclusions when a priori measures are linearly related to the "true" scale values that perfectly reproduce the data.

The method of analysis in Brandstätter et al. (2006) contains an additional problem: They used parameters estimated with one index of fit and then compared models using another index of fit. Whereas parametric models are fit using least squares or maximum likelihood to choice proportions by individuals, heuristic models are devised to maximize percentage correct in "predicting" the modal choices (Brandstätter et al. 2006, p. 425). Least squares solutions to individual choice proportions do not necessarily produce the highest percentage of correct predictions of the modes in aggregate data. Because the priority heuristic does not predict choice proportions or individual data, it does not allow itself to be compared using the same criteria that were used to optimize the parameters of models like CPT and TAX.

One can safely assert that the mode is "better" than the mean because it is more often correct than the mean, but this statement is trivial. For any distribution, the mean minimizes the sum of squared deviations, the median minimizes the sum of absolute deviations, and the mode minimizes the number of errors. For example, if we choose the mode of the numbers 0, 1, 0, 0, 0, 3, 3, 5, 6, it will be correct 44% of the time, and if we choose the mean (2), it will be right 0% of the time. Certainly if we "take the best" value (0), it will be right more often than the mean. However, if we use the mean, our sum of squared errors will be 32, which is less than the sum of squared errors for the mode (80); so by the sum of squares criterion, the mean is "better" than the mode.

So even if we agreed to use percentage correct predicting modal choices as our global index, then we should use that same criterion to estimate parameters for all of the models. Because percentage correct is not a continuous bitonic function of each parameter, use of this criterion creates technical problems in model fitting because this criterion has local minima, nonunique solutions, and "flat" regions where fit is unchanged by small variations in the parameters (e.g., see Figures 1 and 2).

Are these problems of exporting parameters just hypothetical worries that can be ignored? No. Suppose Brandstätter et al. (2006) wanted to use the best-fit parameters of TAX from Birnbaum and Navarrete (1998) in their contests of fit. Should they use the best-fit value to the averaged data ($\beta = 0.73$), or should they use the median best-fit value from individual participants ($\beta = 0.41$)? Figure 1 shows that choosing 0.41 would make TAX appear the better model for the Erev et al. (2002) data, but choosing 0.73 would make priority heuristic seem the (slightly) better model for those data. But neither of those values maximized the number of correct predictions of the modal choices in Birnbaum and Navarrete (1998). That "best" value was $\hat{\beta} = 0.62$ (with $\hat{\gamma} = 0.92$ and $\hat{\delta} = 0.28$, TAX correctly predicted 95 of the 112 choices).

Are there any proper uses for model fitting and parameter estimation at all? Yes. One can use parameter estimates from choices made by a person to predict other choices made in the same context by that same person; indeed, this assumption is Birnbaum's (1982, p. 456) principle of scale convergence (Birnbaum, 1974a, 1983, 1990; Birnbaum & Veit, 1974). When that fails, one can safely reject a model (e.g., see Birnbaum & Sutton, 1992). In addition, one can use parameter estimates from one model to predict where to find critical violations of a rival model.

Testing Critical Properties

Rather than comparing global indices of fit applied to nondiagnostic data using a priori parameter values, a better way to compare models is to explore cases in which the implications are systematically different for rival models. By testing *critical properties*—by which I mean properties that hold true for any functions and any parameters for at least one of the models and not for all of the models—one can make discriminations between theories that are not limited to particular parametric or functional assumptions.

For example, CPT implies first-order stochastic dominance: Despite its free parameters and functions, there is no set of functions or parameters that allows CPT to predict violations of stochastic dominance. Therefore, one does not need to estimate parameters to test CPT in a study of stochastic dominance (e.g., Birnbaum & Navarrete, 1998). Such tests are far stronger than merely showing that with certain parameters, a model does not predict a choice, whereas with other parameters, that same model could predict that same choice.

Birnbaum (2007) devised three new critical properties that can be used to test the family of lexicographic semiorder models. Because the priority heuristic is a type of lexicographic semiorder, it is possible to test a wider set of theories than just the priority heuristic. For example, one can test the possibility that different people might use different priority orders with different threshold parameters. These new tests even allow one to test the theory that the data result from a mixture in which people randomly change

from trial to trial, shifting among different lexicographic semiorders.

Priority Dominance

Priority dominance is the assumption that if a dimension has priority, no change on dimensions with lower priorities should be able to reverse a choice. For example, according to the priority heuristic, people should choose S = (\$50, 0.5; \$49) over R = (\$50, 0.9; \$0) because the lowest consequence of S exceeds the lowest consequence of R by more than 1/10 of the highest consequence in either gamble. If the lowest consequence has priority over the other two dimensions, people should also choose S'' = (\$50, 0.5; \$49) over R' = (\$100, 0.9; \$0) because the lower consequences are the same, and the difference still exceeds 1/10 of the highest consequence. Instead, Birnbaum (2007) found that 81% of 242 undergraduates preferred S over R, but only 35% of the same people preferred S'' over R'. These percentages differ significantly from 50% in opposite directions. Other tests of priority dominance, designed to check if different people might have different priorities, lead to similar conclusions: Very few people appear to obey priority dominance.

Dimension Integration

The test of dimension integration asks if changes on one dimension or another that individually do not reverse a choice can combine to reverse a choice (Birnbaum, 2007). For example, if S = (\$40, 0.6; \$38) is preferred to all three of the following—R = (\$45, 0.1; \$0), R' = (\$100, 0.1; \$0), and R'' = (\$45, 0.9; \$0)—then, as shown next, S should also be preferred to R''' = (\$100, 0.9; \$0), according to any of the priority heuristic models. Note that the difference between R and R' is a change in the highest prize from \$45 to \$100, and the difference between R and R'' is a change in probability from 0.1 to 0.9. If neither of these two changes by itself is enough to reverse the choice, then their combination should not reverse the choice between S and R'''. However, Birnbaum (2007) found that 89%, 77%, and 72% of 266 participants chose S over R, R', and R'', respectively, but only 27% chose S over R''', indicating that people integrate these dimensions. According to the priority heuristic, the majority should have chosen S in every case because the difference in the lowest consequence (\$38 versus \$0) should have been decisive in all four choices.

Brandstätter et al. (2006) concluded that when the ratio of EVs exceeds two, the priority heuristic is not very accurate. Brandstätter et al. (2006) conjectured that either people do, in fact, integrate information or that *decisive differences*—large differences on one dimension or another—were the cause of failures of the priority heuristic that were correlated with EV. The above example of dimension integration includes three choices with ratios of EV outside this range. The example above rules out the decisive difference interpretation of the EV correlation; in other words, the results show that people do in fact integrate. Birnbaum and LaCroix (in press) presented other tests of dimension integration in which the EV ratios fell inside the estimated boundaries of the EV-revised priority heuristic. Those tests also showed clear evidence of integration.

Interactive Independence

Birnbaum (2007) also proposed testing interactive independence, defined as follows:

$$A = (x,p;y) \succ B = (x',p;y') \Leftrightarrow A' = (x,p';y) \succ B' = (x',p';y'),$$

in which Birnbaum found violations when $x \geq x' \geq y' \geq y \geq 0$. Note that the only change between the two choices is that the common probability, p, has been changed to p'. If there is no interaction between probability and consequences, a change in the (common) probability should not reverse the choice between A and B. Instead, Birnbaum (2007) found that 71% of 153 participants chose B = (\$55, 0.1; \$20) over A = (\$95, 0.1; \$5), but only 17% of the same people chose B' = (\$55, 0.99; \$20) over A' = (\$95, 0.99; \$5), which is evidence of dimension interaction.

Transitivity of Preference

A fourth property that can be used to compare the priority heuristic against models like TAX or CPT is transitivity of preference, the assumption that if $A \succ C$ and $C \succ E$, then $A \succ E$. Whereas TAX and CPT satisfy transitivity, the priority heuristic violates transitivity (Brandstätter et al., 2006). For example, the priority heuristic predicts that most people should prefer A = (\$500, 0.29; \$0) over C = (\$450, 0.38; \$0), that they should prefer C over E = (\$400, 0.46; \$0), and that the majority should violate transitivity by choosing E over A because the difference in probability exceeds 0.1 in this choice.

Tversky (1969) reported that transitivity was violated by some people in such choices, but he did not claim that the majority of his participants were intransitive, as implied by the priority heuristic. Birnbaum and Gutierrez (2007) found that 77% of 327 participants were consistent with the transitive order predicted by the TAX model ($E \succ C \succ A$, 22% were consistent with other transitive orders, and only 1% were estimated to be intransitive. Birnbaum (2007), Birnbaum and Gutierrez (2007), and Birnbaum and LaCroix (in press) searched for violations of transitivity predicted by the priority heuristic (including replication of the Tversky study) without finding a single case in which the majority violated transitivity.

EV and the Priority Heuristic

Brandstätter et al. (2006) noted that their model might be made more accurate if it assumed that people compute the EV of each gamble, take their ratio, and choose the gamble with the higher EV if the ratio exceeds two. Suppose people only use the priority heuristic when this ratio is less than two. This modification might be able to save the heuristic from some of the evidence against it because EV is an integrative model with an interaction between probability and consequence.

But even this modification would not account for all of the previous findings. For example, violations of stochastic dominance reported by Birnbaum (1999b, 2004a, 2004b, 2005a) are violations of EV, as well as violations of the priority heuristic. Violations of restricted branch independence and cumulative independence also include cases in which people violate both EV and the priority heuristic. For example, R = (\$97, 0.1; \$11, 0.1; \$2, 0.8) has a higher EV (12.4) than S = (\$52, 0.1; \$48, 0.1; \$2, 0.8), whose EV

is 11.6; however, 68% chose S over R (Birnbaum & Navarrete, 1998), contrary to both EV and the priority heuristic. So even the addition of EV to the list of steps in the priority heuristic does not account for previous results.

Nor does this incorporation of EV into the priority heuristic explain all the new findings by Birnbaum (2007). For example, in several tests of dimension interaction, EV ratios are less than two, and yet people display evidence of interaction. Similarly, the priority heuristic (without the EV modification) predicts that people should prefer the "safe" gamble S = ($51, 0.5; $50, 0.5) over the "risky gamble" R = ($100, 0.1; $50, 0.9)—it has a much lower probability of the lowest consequence and the lowest consequences are equal. Instead, Birnbaum (2007) found that 67% of 242 participants chose the "risky" gamble, which has EV of $55, compared with EV = $50.5 for the "safe" gamble. If we assume that EV is the reason for this violation, we must conclude that the threshold for using EV is less than 9%, in which case most people should prefer ($100, 0.5; $0, 0.5) over $40 for sure. (In this case the risky gamble has an EV of $50, compared with $40, which represents a ratio of 1.25.) But we know that most people prefer $40 in this case, contrary to EV (Birnbaum, 1999b; Tversky & Kahneman, 1992). Furthermore, the gambles used by Birnbaum and Gutierrez (2007) varied by steps as small as 3.3% in EV. If we assume that people did not use the priority heuristic in Birnbaum and Gutierrez (2007) because they used EV instead, we must conclude that people choose by EV ratios as small as 1.033.

These examples show that, although manipulations that affect EV can be used to create violations of the priority heuristic, incorporating EV into the priority heuristic does not provide a consistent account for all the violations of that model. A better way to modify the priority heuristic would be to include a more accurate integrative model, rather than using EV as the first step. For example, one might postulate that people compute TAX and make a decision on the basis of it first. But this approach then leads to the following question: Once we incorporate an accurate integrative model as the first step, are there any phenomena left that require the use of the rest of the priority heuristic for their explanation?

The confession by Brandstätter et al. (2006, p. 426) that gambles similar in EV make "difficult" choices seems an admission that people make decisions by computation and not by a verbal analysis of propositions. But the idea that people use EV, or make any other such computation before deciding to use the priority heuristic, contradicts the theoretical arguments by Brandstätter et al. that people make decisions by verbal heuristics, unless we suppose people are actually making calculations. This admission also seems to contradict the idea that people try to be "fast and frugal" by ignoring some of the information, because calculating EV requires all of the information. Furthermore, EV is an integrative model with trade-offs, whereas Brandstätter et al. (2006) argued that people do not integrate information.

The priority heuristic was intended to provide a "fast and frugal" way for a primitive, language-based decision maker to process risky gambles. If one wanted to be fast, why would he or she first compute EV, take a ratio, and then decide not to use it? Because the EV rule does not account for violations of stochastic dominance, the decision maker has to make two different decisions using all of the information (EV ratio and stochastic dominance) to decide whether to use the priority heuristic.

How much of the domain is excluded by the restriction on EV? Consider again two-branch gambles with prizes uniformly distributed between $0 and $100 and probabilities uniformly distributed between 0 and 1. On the basis of 100,000 simulated random choices, the restriction $1/2 < EV_1/EV_2 < 2$ excluded 39% of such choices. Recall that half this space is ruled out by the stochastic dominance alone; the union of both criteria rules out 59% of the space.

If our domain of interest is all choices between sure cash, c, and (x, p; y)—with uniformly distributed prizes from $0 to $100—the exclusion zone becomes even greater. For 100,000 simulated random pairs, 67% are excluded by stochastic dominance alone, 39% are excluded by EV ratio alone, and only 28% of the choices are not excluded by one or both of these conditions.

Brandstätter et al. (2006) excluded other choices from their model; for example, they conjectured that people would not use the priority heuristic for choices with small consequences. This reduces the space further and requires a third predecision stage that precedes the priority heuristic. The priority heuristic thus applies to a very narrow set of choices, and even within this narrow domain, it does not account for data of Tversky and Kahneman (1992) or Mellers et al. (1992), when these data are properly analyzed. Nor does the priority heuristic account for the results of Birnbaum and Chavez (1997) or Birnbaum and Navarrete (1998), which Brandstätter et al. did not attempt to describe with their model. Finally, even with three extra predecision stages, the priority heuristic does not explain the new results of Birnbaum and Gutierrez (2007), who found that most people satisfy transitivity in cases in which the priority heuristic predicts that most people should violate it. Nor does the priority heuristic account for violations of new critical properties like dimension interaction and dimension integration (Birnbaum, 2007; Birnbaum & LaCroix, in press).

References

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. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp. 401–485). Hillsdale, NJ: Erlbaum.

Birnbaum, M. H. (1983). Scale convergence as a principle for the study of perception. In H.-G. Geissler & V. Sarris (Eds.), *Modern issues in perceptual psychology* (pp. 319–335). Amsterdam: North-Holland.

Birnbaum, M. H. (1990). Scale convergence and psychophysical laws. In H.-G. Geissler, W. Prinz, & M. H. Muller (Eds.), *Psychophysical explorations of mental structures* (pp. 49–57). Toronto, Ontario, Canada: Hogrefe & Huber.

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. (1999a). Paradoxes of Allais, stochastic dominance, and decision weights. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards* (pp. 27–52). Norwell, MA: Kluwer Academic.

Birnbaum, M. H. (1999b). 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. (2006). Evidence against prospect theories in gambles with positive, negative, and mixed consequences. *Journal of Economic Psychology, 27,* 737–761.

Birnbaum, M. H. (2007). *Testing heuristic models of decision making: Priority dominance, dimension integration, and dimension interaction.* Manuscript submitted for publication.

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 semi-order. *Organizational Behavior and Human Decision Processes, 104,* 97–112.

Birnbaum, M. H., & LaCroix, A. R. (in press). Dimension integration: Testing models without trade-offs. *Organizational Behavior and Human Decision Processes.*

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., & Sutton, S. E. (1992). Scale convergence and utility measurement. *Organizational Behavior and Human Decision Processes, 52,* 183–215.

Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception and Psychophysics, 15,* 7–15.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Choices without tradeoffs. *Psychological Review, 113,* 409–432.

Erev, I., Roth, A. E., Slonim, R. L., & Barron, G. (2002). *Combining a theoretical prediction with experimental evidence to yield a new prediction: An experimental design with a random sample of tasks.* Unpublished manuscript, Columbia University and Faculty of Industrial Engineering and Management, Techion, Haifa, Israel.

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

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.

Mellers, B. A., Chang, S., Birnbaum, M. H., & Ordóñez, L. (1992). Preferences, prices, and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 347–361.

Rieger, M. O., & Wang, M. (in press). What is behind the priority heuristic: A mathematical analysis and comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, XX,* xxx–xxx.

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.

Received January 31, 2007
Revision received March 13, 2007
Accepted March 16, 2007 ■

Postscript: Rejoinder to Brandstätter et al. (2008)

Michael Birnbaum
California State University, Fullerton

Brandstätter, Gigerenzer, and Hertwig (2008, Figure 1) reanalyzed data of Erev, Roth, Slonim, and Barron (2002) to display a correlation between the accuracy of the priority heuristic and expected value (EV). From this correlation, they argued that people use two processes to choose between gambles: First, they act as if they compute ratios of EV and choose the gamble with the higher EV when this ratio exceeds a threshold, which the authors estimated to be two. Given this parameter, 56 of the 100 choices of Erev et al. were supposedly decided by EV. Second, people supposedly used the priority heuristic to make the remaining 44 choices. Whereas the priority heuristic alone had 15 errors predicting these 100 modal choices, the EV rule plus priority heuristic had six errors. But even with EV, a free parameter, and a lexicographic heuristic, this model does not fit as well as either the transfer of attention exchange (TAX) model or cumulative prospect theory (CPT), each of which had only one error when the same choices were reproduced. Because several models can fit these data almost equally well, I think it best to say that these data, like the Kahneman and Tversky (1979) data, are simply not diagnostic for comparing those models. Given the fact that they use an "as-if" (EV) model and estimate its parameter (two) post hoc, it seems odd that Brandstätter et al. continue to argue against "as-if" models and parameter estimation.

The EV plus priority heuristic model does not fit other, more diagnostic studies, as Brandstätter et al. (2008) acknowledged. To handle such data, they argued that each new failure of the EV plus priority heuristic should be taken as evidence for another heuristic. Among the additional heuristics they added to the theoretical stew are "dominance," "similarity," "toting up," "cancellation," "combination," and the "most-likely" heuristic. Consider how their "toting-up" heuristic was devised. Because EV plus priority heuristic is correct for fewer than half of the modal choices in Birnbaum and Navarrete (1998), Brandstätter et al. decided to replicate part of that study. There were 144 choices in the original study, including 112 in the main experimental designs that tested stochastic dominance and cumulative independence. For reasons I do not understand, Brandstätter et al. decided to examine only 54 of those 112 choices, so their new data did not allow tests of stochastic dominance or of upper or lower cumulative independence. Also puzzling were their decision to use a smaller number of participants than in the original study, their use of two new formats for the presentation of gambles, and other changes in procedure. Unlike the study by Birnbaum and Navarrete, every