frontiers in
PSYCHOLOGY

OPINION ARTICLE
published: 15 November 2011
doi: 10.3389/fpsyg.2011.00315

Testing theories of risky decision making via critical tests

Michael H. Birnbaum*

Department of Psychology, California State University, Fullerton, CA, USA
**Correspondence: mbirnbaum@fullerton.edu*

Whereas some people regard models of risky decision making as if they were statistical summaries of data collected for some other purpose, I think of models as theories that can be tested by experiments. I argue that comparing theories by means of global indices of fit is not a fruitful way to evaluate theories of risky decision making. I argue instead for experimental science. That is, test critical properties, which are theorems of one model that are violated by a rival model. Recent studies illustrate how conclusions based on fit can be overturned by critical tests.

Elsewhere, I have warned against drawing theoretical conclusions from indices of fit (Birnbaum, 1973, 1974, 2008a): Fit changes under monotonic transformation of the dependent variable and scaling of stimuli. An index of fit depends on experimental design; it depends on parameters and how they are estimated. Different indices can lead to opposite conclusions. A wrong model can achieve a "good" fit, and it can even fit better than the model used to generate the data. I will not add here to this list of problems; instead, I argue in support of traditional science.

A theory is a set of statements satisfying five philosophical criteria: (1) it is *deductive* in that the phenomena to be explained can be derived from the theory; (2) it is *meaningful*; that is, it can be tested (potentially falsified); (3) *predictive*: if we knew the theory, in principle, we could have predicted the events to be explained; (4) *causal*: it specifies in principle how to alter the phenomena via manipulation; and (5) *general*: premises used in a theory are *laws*; they are not assumed or denied from case to case.

In deduction, when premises are true, conclusions must also be true. However, if a conclusion is assumed (or empirically established), it says nothing about the truth of the premises. Therefore, we cannot "prove" a theory via experiments. However, if implications deduced from a theory are false, we know the theory is false. So we can test a theory by testing its theorems. A test is an opportunity to disprove, but failure to disprove does not prove a theory.

The term "model" refers to a special case of a theory that also includes all of the operational definitions and simplifying assumptions needed to apply a theory to a particular paradigm.

The classic paradoxes of Allais (1953) are examples of critical tests. These paradoxes lead expected utility (EU) theory into self-contradiction. They do not require us to estimate any parameters from data, nor do we need to compute an index of fit, because the "paradoxical" behavior, if real, shows that no parameters will work. Models proposed to account for these paradoxes include prospect theory (PT; Kahneman and Tversky, 1979), cumulative prospect theory (CPT) (Tversky and Kahneman, 1992), and the transfer of attention exchange (TAX) model (Birnbaum, 1999).

Because people often make different responses when the same choice problem is repeated, it is useful to distinguish instability of preference due to random error from that due to a false theory. The true and error model assumes that different people may have different "true" preferences when presented with a given choice problem, that different choice problems may have different error rates, and that some individuals may have more "noise" in their data than others (Birnbaum, 2008c, Appendix D; Birnbaum and Gutierrez, 2007). This model provides a neutral standard for testing critical properties, such as Allais paradoxes and new paradoxes that distinguish between CPT and TAX.

The original version of PT had a number of problems that required a list of "editing rules," added to excuse the model from potential evidence against it. For example, PT implied that people would violate stochastic dominance in cases where all possible consequences of one gamble are better than the best consequence of the other. So a rule was added to say that people satisfy dominance whenever they detect it, but it did not say when people detect it. CPT solved this problem, because it implies that people always satisfy stochastic dominance, apart from random error.

A configural weighting model (Birnbaum and Stegner, 1979), implies that dominance is not always satisfied. A simple version of this model was fit to risky decisions, where it was renamed the TAX model, and a recipe was constructed for choices in which the model predicts a violation (Birnbaum, 1997). Here is an example:

Urn *A* contains:	85 Tickets to win \$96
	5 Tickets to win \$90
	10 Tickets to win \$12
Urn *B* contains:	90 Tickets to win \$96
	5 Tickets to win \$14
	5 Tickets to win \$12

One ticket will be drawn randomly from the chosen urn, to determine the prize. Which urn would you choose? According to CPT, people should prefer *B*. One need not estimate any parameters, because CPT makes this prediction for any set of parameters and any monotonic value and probability weighting functions. Although TAX can satisfy stochastic dominance (EU is a special case of TAX), it violates dominance in this choice for plausible parameters (Birnbaum and Navarrete, 1998; Birnbaum, 2004a, 2005, 2008b).

A critical property is a theorem of one theory that is violated by a rival. In this case, CPT with any parameters implies people must choose *B* (apart from random error), but TAX with parameters predicts *A*. Such choices have now been tested with thousands of people, using a dozen formats for presenting choices. About 60–70% of undergraduates violate CPT by choosing *A* instead of *B*, contrary to stochastic dominance, in a single choice of this type. When corrected for unreliability of responses, the estimated rate of "true" violation is even higher (Birnbaum, 2004b, 2008b, Table 11).

According to the TAX model, the utility of the gamble is a weighted average of the utilities of the consequences, with

weights that depend on probability and on the ranks of the consequences. Because the weighting function for probability is negatively accelerated, a branch with five tickets (0.05) ends up getting relatively more weight compared to its objective probability, which causes *A* to appear better because the 0.05 branch to win \$90 in *A* (and the 0.05 branch to win only \$14 in *B*) get more weight.

Other critical tests also refute CPT. Empirical studies of 12 theorems of CPT show that neither version of PT can be retained as descriptive of risky decision making (Birnbaum, 2008b,c).

Brandstätter et al. (2006) proposed the priority heuristic (PH) based on an index of fit assessing how this model performed in describing the data used to generate the model. The PH is a variant of a lexicographic semiorder (LS) used by Tversky (1969) to describe violations of transitivity. They claimed PH was more often correct in predicting modal choices than either CPT or TAX, both of which are transitive models. But these conclusions reverse when parameters are estimated instead of fixed in advance; they reverse when we consider different sets of data, and most important: they reverse when we examine critical properties designed to test these theories.

The family of LS, including PH, must satisfy interactive independence. People should make the same decisions in these two choices:

Choice 1:	
Urn *C* contains	90 tickets to win \$100
	10 tickets to win \$5
Urn *D* contains	90 tickets to win \$50
	10 tickets to win \$20
Choice 2:	
Urn *E* contains:	10 tickets to win \$100
	90 tickets to win \$5
Urn *F* contains:	10 tickets to win \$50
	90 tickets to win \$20

According to PH, people should choose *D* (over *C*) and *F* (over *E*) because the lowest consequence is better and the difference (\$15) exceeds threshold. According to any member of the LS family (with different orders of examining the attributes, different psychophysical functions on the attributes, and different thresholds) a person should either choose *C* and *E* or *D* and *F*, or be indifferent in both, but she should not switch, except by error, because any attribute that is the same in both alternatives (here probability is the same) should have no effect. Instead, the true and error model indicated that 63% of those tested switched their true preferences from *C* to *F* (after correcting for preference instability due to random error), demonstrating an interaction between probability and the prizes (Birnbaum, 2008c).

Other critical tests also refute LS and PH (Birnbaum, 2008c, 2010). PH may have looked "good" by means of an index of fit applied to certain studies using fixed parameters, but it has not been successful in predicting new results.

If a critical test is satisfied, it does not mean that the theory that implies it is "validated," "confirmed," or "proved." It merely means that the theory that implies it can be retained. However, the greater the number of interesting predictions that a theory makes that are satisfied, the more we are likely to bet on its predictions in the future. Thus, confidence in a theory can grow by induction, but scientific theories are always open to revision or refutation based on new evidence.

Does testing theories via critical properties mean that there is no role for model-fitting and parameter estimation? No. These serve two important functions: First, we should try to learn from our data where a model fits poorly, in order to devise new tests that have the potential to refute the model. Second, parameters are used to devise new tests between rival models.

For example, PH was devised to account for previously published data, such as those of Tversky (1969) who reported violations of transitivity consistent with a LS (Brandstätter, et al., 2006, 2008). Transitivity is the assumption that if *A* is preferred to *B* and *B* is preferred to *C*, then *A* should be preferred to *C*. Because PH can account for violations of transitivity and models like EU, CPT, and TAX cannot, transitivity is a critical property that has the potential to refute both CPT and TAX.

Just as the TAX model had been used to construct a test of stochastic dominance where violations of CPT should be observed, PH has been used to design new tests of transitivity to search for predicted violations of TAX and CPT that satisfy this critical property.

Birnbaum and Gutierrez (2007) and Regenwetter et al. (2010, 2011) carried out such tests, using designs similar to those of Tversky (1969), but they were not able to find much, if any, evidence for the predicted intransitive behavior. Birnbaum and Bahra (2007) devised three interlaced designs in which PH predicted violations of transitivity. Although they found evidence that perhaps as many as 4% of participants were partly or momentarily intransitive, they were not able to refute transitivity for the vast majority of cases. The PH was correct in predicting modal choices in only 18 of 60 new choices devised to test its predictions (30%).

This case illustrates how conclusions based on an index of fit can be ephemeral. What looks good by an index applied to selected data can look horrible when that model and its parameters are used to predict the results of a new study testing critical properties.

REFERENCES

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica* 21, 503–546.

Birnbaum, M. H. (1973). The devil rides again: correlation as an index of fit. *Psychol. Bull.* 79, 239–242.

Birnbaum, M. H. (1974). Reply to the devil's advocates: don't confound model testing and measurement. *Psychol. Bull.* 81, 854–859.

Birnbaum, M. H. (1997). "Violations of monotonicity in judgment and decision making," in *Choice, Decision, and Measurement: Essays in Honor of R. Duncan Luce*, ed. A. A. J. Marley (Mahwah, NJ: Erlbaum), 73–100.

Birnbaum, M. H. (1999). "Paradoxes of Allais, stochastic dominance, and decision weights," in *Decision Science and Technology: Reflections on the Contributions of Ward Edwards*, eds J. Shanteau, B. A. Mellers, and D. A. Schum (Norwell, MA: Kluwer Academic Publishers), 27–52.

Birnbaum, M. H. (2004a). Causes of Allais common consequence paradoxes: an experimental dissection. *J. Math. Psychol.* 48, 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. *Organ. Behav. Hum. Decis. Process* 95, 40–65.

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

Birnbaum, M. H. (2008a). Evaluation of the priority heuristic as a descriptive model of risky decision making: comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychol. Rev.* 115, 253–262.

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

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

Birnbaum, M. H. (2010). Testing lexicographic semi-orders as models of decision making: priority dominance, integration, interaction, and transitivity. *J. Math. Psychol.* 54, 363–386.

Birnbaum, M. H., and Bahra, J. P. (2007). Transitivity of preference in individuals. *Society for Mathematical Psychology Meetings*, Costa Mesa, CA.

Birnbaum, M. H., and Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semiorder. *Organ. Behav. Hum. Decis. Process* 104, 97–112.

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

Birnbaum, M. H., and Stegner, S. E. (1979). Source credibility in social judgment: bias, expertise, and the judge's point of view. *J. Pers. Soc. Psychol.* 37, 48–74.

Brandstätter, E., Gigerenzer, G., and Hertwig, R. (2006). The priority heuristic: choices without tradeoffs. *Psychol. Rev.* 113, 409–432.

Brandstätter, E., Gigerenzer, G., and Hertwig, R. (2008). Postscript: rejoinder to Johnson et al. (2008) and Birnbaum (2008). *Psychol. Rev.* 115, 289–290.

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

Regenwetter, M., Dana, J., and Davis-Stober, C. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Front. Psychol.* 1:148. doi: 10.3389/fpsyg.2010.00148

Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychol. Rev.* 118, 42–56.

Tversky, A. (1969). Intransitivity of preferences. *Psychol. Rev.* 76, 31–48.

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain* 5, 297–323.

Received: 01 August 2011; accepted: 17 October 2011; published online: 15 November 2011.

Citation: Birnbaum M (2011) Testing theories of risky decision making via critical tests. Front. Psychology ***2****:315. doi: 10.3389/fpsyg.2011.00315*

This article was submitted to Frontiers in Cognition, a specialty of Frontiers in Psychology.