How to Use a Multi-Criteria Comparison Procedure to Improve Modeling Competitions.

Jason L. Harman[1]*, Michael Yu[2], Emmanouil Konstantinidis[3], & Cleotilde Gonzalez[2]

[1]Louisiana State University
[2]Carnegie Mellon University
[3]University of Warwick

Accepted for publication in *The Psychological Review* Nov. 14, 2020

Keywords: Choice Prediction Competition; Modeling Competitions; Multi-Criteria Competitions; Cognitive Modeling

Author Note: The Authors would like to thank Dr. Nicholas Mattei and two anonymous reviewers for their helpful suggestions in the final version of this work.nAll the authors were part of the Dynamic Decision Making Laboratory (www.cmu.edu/ddmlab) at Carnegie Mellon University, when this work was performed. They have since then, moved to different universities. Some of the ideas presented in this paper were presented in an early form at the Society for Judgment and Decision Making (Harman, Yu, Morrison, Konstantinidis, & Gonzalez, 2017) and the Oklahoma Kansas Meeting of Judgments and Decision Making (Harman, 2017). The research was supported by the National Science Foundation Award #1530479.

Abstract

Modeling competitions are a promising method for advancing psychological science. In this commentary to Erev, Ert, Plonsky, Cohen, & Cohen (2017), we highlight how this promise could be enhanced through modifying competition structures to produce insights more directly in line with the goals of promoting psychological knowledge. We argue that a single criterion on which models are compared limits the diversity of models entered into competitions, restricting the number and type of insights that can be gained consequently. We propose an alternative competition structure with multiple evaluative criteria and outline a quantitative selection method for choosing a winner. Our proposed competition structure has the advantages of: a) increasing the diversity of models entered, b) incentivizing desirable qualities of models, c) disambiguating competition winners, and d) enhancing the impact and possible insights gained from competitions, all these while allowing flexibility for competition organizers to emphasize specific qualities of models.

Introduction

Modeling competitions have become a popular research method in many scientific fields. Competitions have shown to be an effective methodology for maximizing the benefit of large sets of data by crowdsourcing modeling to multiple research groups across multiple research fields. In computer science, examples of modeling competitions have ranged from recommendation systems (Bennett & Lanning, 2007), and bioinformatics (Triguero et al., 2015; Goldbloom, 2010) to data mining (Weigend & Grenshenfeld, 1993) and machine learning (Sajda, et al. 2003). In fact, multiple organizations such as Drivendata (drivendata.org) and Kaggel (kaggel.com) have been formed to utilize modeling competitions to advance solutions to real-world problems such as disease spread, clean drinking water, and disaster risk. This methodology has been increasingly promoted as a means of reconciling and refining psychological theories (Erev, et al., 2010; Lai, et al., 2014; Gonzalez & Dutt, 2011; Lebiere et al., 2010). In psychological literature, variability in data collection methods, sample characteristics, and theoretical motivations can produce contradictory examples and predictions of behavioral effects (Erev, Ert, Plonsky, Cohen, & Cohen, 2017). Modeling competitions help resolve such contradictions by bringing competing theories together under a unified experimental paradigm.

The goals of psychological research however do not always align with the goals of other fields, and modeling competitions in psychology may be disserviced by uniformly mimicking all aspects of competitions from computer science. While machine learning, for example, may place a stronger emphasis on the accuracy of predictions, psychological research (in addition to the paramount goal of improving predictions) also seeks to develop deeper insights into the processes underlying human behavior. This stems from the broader goal in psychological research to develop and test general theories of human behavior. Descriptive theories of behavior postulate processes and mechanisms that govern general phenomena. But it is the quantitative nature of a theory that can make it precise and testable (Gonzalez, 2017; Gonzalez, Lerch, & Lebiere, 2003). To test theories of human behavior, we use computational models: representations of some or all aspects of a theory as it applies to a particular task or context. Thus, the value of models is that they can represent concrete problems and provide explicit mathematical and computational representations of a theory, which can then be used to make predictions about behavior.

A model that can capture and explain the underlying mental processes proposed by a theory, can be applied to predictions of similar behavior in different contexts, with different actors, and at different time points. In this paper, we argue that modeling competitions have great potential to advance psychological theory and practice, but that the structure of competitions in psychology can better enable this potential by evaluating and incentivizing more than a single prediction criterion. Specifically, we use a recent decision making modeling competition (Erev, Ert, Plonsky, Cohen, & Cohen, 2017) to illustrate how using a single criterion to evaluate models may limit the diversity of models entered, and consequently limit the insights that could be gained from the competition. We also propose a method of *competition design* that would mitigate these limitations in future competitions and offer multiple advantages.

There are many reasons to like modeling competitions, including rapid and efficient advancement using large data sets and providing uniform comparisons between different models. This is especially true in the modern world where large data sets, computational power, and coordination between multiple researchers are increasingly manageable. First, crowdsourcing science is an efficient way to quickly advance a field. The Netflix prize awarded $1 million to the

team that won a prediction competition improving movie recommendations (netflixprize.com). Though that seems like a large amount of money, it prompted many thousands of work hours from some of the best computer scientists in the world and financially was a bargain (Hunt, 2014).[1] Second, competitions can lead to rapid improvements on very specific problems. In a bioinformatics modeling competition, competing models had to pick genetic markers in HIV sequences that correlate with viral load (i.e., severity). Within a week and a half entrants had already outperformed the best methods in the scientific literature (Goldbloom, 2010). Third, competitions allow for direct model comparisons in identical scenarios. Traditional modeling exercises and comparisons can become both difficult and time consuming to evaluate, as papers vary in scope, stimuli, data, modeling frameworks and evaluation rules (e.g., González-Vallejo, Harman, Mullet & Muñoz-Sastre, 2012). Competitions overcome this by making every entrant compete on comparable tests. In line with standardizing comparisons, competitions allow for the testing of multiple auxiliary hypotheses between models beyond the criteria set out in the competition. For example, Lai, et al. (2014) hosted a competition to test different methods of reducing implicit racial bias. While no intervention reliably reduced implicit bias, comparison of the 17 different interventions did show that interventions involving some sort of counter typical exemplars and conditioning were *less* ineffective than higher order interventions like perspective taking and considering egalitarian views.

With all of the possible benefits of modeling competitions for psychology, it is important to consider the goals of modeling competitions and how those goals can be best accomplished. There has been previous work defining the advantages and disadvantages of different competition structures and statistical criteria of model selection (e.g., Spiliopoulos & Ortmann, 2014). Spiliopoulos and Ortmann (2014) provide an extensive overview of types of modeling competitions noting among other things that the results of competitions may be sensitive to the exact criterion chosen and that a side effect of using a uniform paradigm is a lack of generalizability. Here we focus on the specific question of how the evaluation criteria affects 1) the models entered in competition and 2) the subsequent insights gained from competitions. To date, most modeling competitions in psychology compare models along one criterion, predictive accuracy, using a single metric or statistical index of goodness of fit such as mean squared deviation (MSD). Our primary argument is that in psychology, multiple competition criteria would increase both the diversity of models entered as well as the number of insights that could be gained from a single competition. Additionally, we outline a quantitative method for comparing models across multiple different criteria in the context of a competition.

To illustrate our arguments we use the Choice Prediction Competition 2015 (CPC2015; Erev et al., 2017), a recent competition of a series of modeling competitions by Erev and colleagues (Erev et al, 2010, Erev, Ert & Roth, 2010, Plonsky, et al. 2019). We use the CPC2015 to illustrate our arguments in order to improve future modeling competitions. We believe the CPC2015 was a major accomplishment which has and will continue to contribute to the understanding of the psychology of decision making. As direct participants of the CPC2015, we are able to retrospectively identify how the competition could have been enhanced. In the following section we outline the CPC2015, discuss the main results, and identify potential missed opportunities that resulted from the use of a single evaluative criterion. We then outline desirable qualities and characteristics of psychological computational models and present a formal method of constructing

[1] Interestingly, the winning algorithm was never used by Netflix as it was too complex to implement effectively (https://medium.com/netflix-techblog/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429).

and running future competitions with multiple evaluative criteria. Finally, we discuss the benefits of the proposed methodology for modeling competitions.

Summary of the CPC2015

The motivation of the CPC2015 was the proliferation of choice anomalies in the literature of decision making under risk, with no real progress towards unifying models that could account for choice behavior across different contexts. The field of decision science has excelled at creating models to explain numerous choice anomalies and choice in differing contexts, but these models tend to be conceptually disconnected from one another leading to an inability to reliably model choice in novel situations or account for multiple choice behaviors within one framework. In many ways, this state of affairs in decision science resembles that of Psychology many years ago with a science driven by exploring numerous individual phenomena extensively while limiting the progress of our understanding of the mind (Newell, 1973), and the suggestions of building complete process models that can improve the level of task integration.

A notable example of this discontinuity in risky decision-making research is the lack of integration of the underlying processes of decision making in descriptive gambles - choices where the options, outcomes, and associated probabilities are given before making a choice (accounted for by Prospect Theory and its many successors and functional adaptations; e.g., Kahneman & Tversky, 1979; Wakker, 2010; Tversky & Fox, 1995), and decision making in experiential choice - repeated choices between options where outcomes and probabilities are not necessarily given beforehand but learned through experience (accounted for by many models Hertwig, 2015; e.g., Instance-Based Learning Theory: Gonzalez, et al., 2003; Gonzalez & Dutt, 2011). Differences in observed choices for similar gambles when presented as one-time descriptive choices or repeated experiential choices has been labeled the *description-experience gap* (Hertwig, Barron, Weber & Erev, 2004; Hertwig & Erev, 2009). The concept of a description-experience gap also led to a dichotomy of models that either captured descriptive-based choices or experience-based choices, but not both (see related arguments in Gonzalez & Dutt, 2011).

The proliferation of decision making models that account for limited anomalies or contexts, exemplified by the description-experience gap, motivated the creation of the CPC2015 where models would be required to account for 14 well known choice anomalies, specific choice scenarios where median choice is contrary to predictions of rational economic theory (e.g., the Allais Paradox, Allais, 1953), and be tested against competing models in a new unique data set involving descriptive and experience-based choices. To accomplish the goals of the CPC2015, Erev and colleagues created an experimental paradigm that could replicate known choice anomalies across different contexts. The paradigm is a 25 trial repeated choice task that gives participants a description of the options they have to choose between (i.e., outcomes and probabilities are stated; see figure 1). In the first five trials, participants do not receive feedback about the outcomes of their choices (i.e., they make decisions from description, DFD). In trials 6-25 participants receive feedback about the outcome of their choice on each trial (i.e., they make decisions from experience, DFE).[2] Using this paradigm Erev et al. created 30 choice pairs which together replicated 14 well-known choice anomalies. An additional 60 problems were created to

[2] Additional modifications were added to this paradigm to replicate decision anomalies such as ambiguity aversion and the St. Petersburg paradox (see Erev et al., 2017 for full paradigm details). We have omitted these details from our description for brevity.

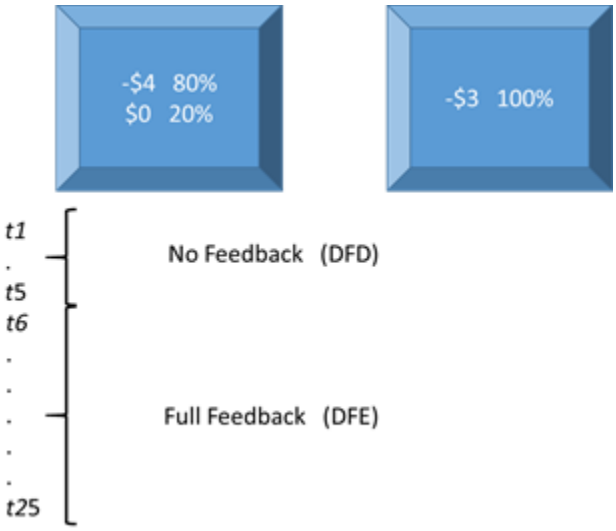
complete the calibration data set (data made available to fit models) and another 60 problems for the prediction data set (data held back to test model predictions).

Although there are common methods inherited from the earlier competitions run by Erev and colleagues (Erev et al., 2017, Erev et al., 2010, Erev, Ert & Roth, 2010, Plonsky et al., 2018)., the CPC2015 was unique in several respects. In previous competitions, Erev and colleagues created a stimuli set relevant to the domain being evaluated, collected data for use by entrants to fit prospective models, provided a baseline model that would be used as a benchmark, and evaluated models based on a single statistical index: the Mean Squared Deviation (MSD) between model predictions and a new prediction dataset (with the same characteristics as the calibration dataset). For the CPC2015, organizers first replicated 14 well-known choice anomalies (10 in description-based choice and 4 in experience-based choice; see Erev et al., 2017 for details) with a single experimental paradigm. The organizers developed a single baseline model (*Best Estimate And Sampling Tools,* BEAST), that was able to capture all behavioral effects in both descriptive and experiential choice. Then, they sent an open invitation for researchers to develop computational models that would do better than BEAST provided they could account for the 14 anomalies using a provided data set generated through human experiments. Those models that captured the anomalies were then submitted to a competition, in which the lowest MSD against a new experimentally-collected data set determined the winner.

In total the CPC2015 consisted of 14 classic choice anomalies embedded in data from a calibration set of 90 decision problems provided to researchers to fit their models, and data from a prediction set of 60 problems that was withheld from researchers until after the competition was completed.

Figure 1

Illustration of the choice paradigm from the CPC2015



Note. Participants are presented with two options that include descriptions of the possible outcomes and associated probabilities and make 25 repeated choices between the options. The first 5 choices produce no feedback (DFD) while the final 20 choices give full feedback (DFE).

Features, results and critiques

The primary measure on which models were compared was the MSD between model predictions and the prediction data set. The distinction between calibration ("fit") and prediction datasets used by CPC2015 is a method of penalizing models that are over-fit, though there are arguments for alternative methods of comparing model complexity and generalizability (see Gonzalez, Dutt & Lejarraga, 2011; Gonzalez & Dutt, 2011). Other criteria were part of the competition, though they did not serve as comparisons between models. A predominant feature of the CPC2015 was the requirement that any qualifying model had to account for all 14 classic choice anomalies. This requirement is in line with reproductive power and scope (discussed more in subsequent sections), however we will argue that the use of this criterion as a gate to entry unduly harmed the diversity of models entered. A final aspect of the competition entry was the requirement for entries to submit both code for their model as well as a verbal description which would be programmed by a third party based only on the verbal description.

Quantitatively, 53 research groups registered for the competition before the deadline for registration and 25 of these groups submitted a final model (Erev et al., 2017). With one exception, the models that were entered into the CPC2015 could be classified as one of three types; 1) minor extensions of the baseline model provided, 2) variants of prospect theory, and 3) statistical or machine learning algorithms. Similar to previous competitions (Erev et al, 2010, Erev, Ert & Roth, 2010), all of the top performing models were minor variants of the baseline model. The machine learning/statistical models uniformly performed poorly in the competition (i.e., fitting the first dataset near perfectly while predicting new data poorly), while the prospect theory variants finished in the middle of the pack. The one exception, described but not named by Erev et al. was the Lexicographic Instance Based Learning Model (LIBL). We developed LIBL not with the goal of minimizing MSD, but with the goal of creating a psychologically plausible process model of decision making that integrated DFD and DFE while meeting the criteria of competition entry[3]. The fact that only one entrant that deviated from either the baseline, prospect theory or machine learning[4] reinforces the idea that the structure of the competition led to a homogeneous group of models entered.

The main insights gained from the CPC2015 were that 1) BEAST predicted new data well using a combination of expected value and four psychological tendencies captured in the sampling tools (pessimism, equal weighting bias, payoff sign sensitivity, and regret minimization), 2) variants of Prospect Theory were able to account for known anomalies, but were not competitive with variants of BEAST in predicting new data, and 3) machine learning models, which fit the calibration data near perfectly, performed the worst in predicting new data. Perhaps most promisingly moving forward, results of the CPC2015 prompted investigation into combining behavioral models (BEAST) and machine learning models (Plonsky, Erev, Hazan, & Tennenholtz, 2017; Plonsk, et al, 2019; Bourgin, Peterson, Reichman, Russell, & Griffiths, 2019). Based on the performance of both machine learning models as well as BEAST, and using data from the CPC2015, Plonsky and colleagues (as well as Bourgin, et al., 2019) have shown that using behavioral decision models to inform the features used in machine learning algorithms can

[3] We also submitted a second model, combining BEAST (1st 5 trials) and IBL (last 20 trials). This model (BIBL) finished second in the competition and was statistically indistinguishable from the winning model.

[4] Arguably, this was also the only model entered that was primarily designed to account for the data with identifiable psychological processes.

improve predictive performance markedly. This idea has prompted an additional modeling competition (Plonsky et al., 2018).

This final insight from the CPC2015 is fundamentally different from the first three in that the ability of descriptive models to improve machine learning models was not designed into the competition but was made possible by comparing the performance of the two different types of models on one criterion. It is this type of unplanned insight that we believe could be enhanced in future competitions. Whereas the CPC2015 wound up with three types of models compared on one criterion, designing competitions with multiple comparative criteria in a way that promotes a diverse set of model entrants would enable a much larger set of auxiliary hypotheses and ideas to be tested.

Models with differing assumptions being compared on the same data allow better exploration of these different assumptions. In a competition like the CPC2015, the number of auxiliary hypotheses that could be evaluated is limited by the diversity of the models entered. With a limited set of models and underlying model assumptions the number of hypotheses that could be evaluated (and in turn the possible insights gained from the competition) was also limited. We believe that part of this lack of diversity stemmed from both the use of MSD as the single criterion for comparison (incentivizing models built with the single goal of prediction while not incentivizing other desirable qualities of a good model) and the use of other criteria (i.e., accounting for classic anomalies) as exclusionary variables.

The use of generalizability (accounting for all 14 choice anomalies) as an exclusionary criterion eliminated generalizability insights. One of the goals of the CPC2015 was to promote models that can capture behavior in a diverse set of circumstances, but without comparison criteria for generalizability, no insight on how anomalies are captured or whether one method is better than another are possible. Of the 28 groups who registered for the competition but did not submit a model, many were for unrelated reasons but at least two were groups that could not account for all 14 anomalies (personal communication). Additionally, some anomalies were not diagnostic with very small effects observed (one anomaly in fact - the reflection effect - was not replicated in the competition data, though models were required to account for the effect). It is possible that the exclusionary structure of the 14 anomaly requirement prevented models from entering the competition on spurious grounds, and that these models could have enhanced the level of insights gained from the competition generally. Other forms of insights (such as identifiable process assumptions) were neither measured nor incentivized.

To remedy some of these shortcomings we suggest an alternative competition structure with multiple criteria, all used for comparison. Multiple comparative criteria (i.e., prediction, generalizability, reproductive power, parsimony, etc.) without exclusionary criteria would promote a more diverse group of model entrants, increasing the number of implicit hypothesis tests and enhancing the insights that could be gained from a competition. Picking a winner based on multiple, not necessarily comparable, criteria seems a difficult task. We propose and discuss one possible method of selection based on social choice literature and present an argument for its superiority to a single criterion selection rule. The only requirement for this selection method is an ordinal ranking (including binary rankings) of models for each criteria. In the following sections, we discuss desirable qualities a good model should have followed by an outline of how multiple criteria could be quantifiably instantiated in this system along with an example of how the selection of a winner would be carried out.

What makes a good model

Before going into detail about competition criteria and how to improve the use and insights gained by such competitions in psychological and cognitive sciences, it is imperative to define the characteristics of a "good" model. To frame what we consider a good model, we outline three general criteria that are important to psychological models along with specific aspects of those criteria (partly adapted from Meir, Lev, & Rosenschein, 2014). We elaborate on these criteria in the subsequent section. It is important to note that models vary widely in their scope and purpose. Additionally, competitions may vary in their goals and theoretical perspective, making some of the following criteria more or less important. While some of the criteria we outline here would not apply to certain models or competitions, we believe that setting a foundation of what models can provide is an important starting point. Following our outline of qualities desirable in a model, we provide suggestions for quantitative measures of each criterion and a method for combining multiple criteria into a single mechanism to select winners. We conclude by discussing how this new method would avoid some of the shortcomings of past tournaments and would lead to future competitions that could provide more insights and advancements to psychology as a whole.

1. Theoretical criteria

1.1 Intuitive understanding- A model should be able to guide intuitive predictions and interventions/prescriptives in the real world. (see Katsikopoulos, 2020; 2014)

1.2 Broad scope- a model should be able to be applied to (or easily adapted to) various scenarios / paradigms. (see Busemeyer & Wang, 2000)

2. Psychological Criteria

2.1 Realistic knowledge- Predicted behavior should not be based on information participants are not likely to have, or is hard to obtain. (Meir, Lev, & Rosenschein, 2014)

2.2 Realistic capabilities- Predicted behavior should not rely on complex computations, non-trivial probabilistic reasoning, etc. (Busemeyer & Diederich, 2010)

2.3 Identifiable process assumptions- A model should rely on identifiable and testable psychological processes. (Weber & Johnson, 2012)

3. Scientific Criteria

3.1 *Parsimony*- a model should have as few parameters as possible, and parameters should be meaningful. (Kuhn, 1977)

3.2 Predictive power / validation - A model should be able to predict new behavioral data with accuracy. (Busemeyer & Wang, 2000)

3.3 Reproductive Power - a model should be able to reproduce common phenomena. (Erev et al., 2017)

3.4 Testability / Falsifiability- A model should produce predictions that could be falsified, or predict behavior that would not happen. (Popper, 1934/1959; Roberts & Pashler, 2000)

Though we have tried to be comprehensive, the criteria and specific aspects of models listed above may be incomplete or may use different language than other works on modeling and model comparisons. Nonetheless, it is a useful guide for the following section discussing insights that could be gained from modeling competitions and the types of evaluation criteria possible. As outlined in the previous section, the CPC2015 was primarily interested in prediction (*3.2*), but also contained aspects relating to replication (*3.3*), parsimony (*3.1*), scope (*1.2*), and intuitive understanding (*1.1*). Only prediction though was used as a comparative quantitative measure,

while the others were exclusionary criteria. In the next section we discuss each criterion in more depth, outlining how each could be instantiated as a comparative criterion in future competitions.

How a Multi-Criteria Competition Could be Established

Theoretical Criteria

As outlined previously, one dimension of a good model is the theoretical insights a model can provide. Two important measures of the theoretical soundness of a model include (*1.1*) *intuitive understanding*: how a model can guide understandable predictions or interventions in real world problems, and (*1.2*) *broad scope*: the ability of the model to be applied (or adapted) to varying scenarios. The benefits of intuitive understanding can best be illustrated by behavioral insights teams or "nudge units" constructed by several governments recently, which use decision-making models such as Prospect Theory to help policy makers derive more accurate intuitive predictions of the impact of different changes in the structure or enforcement of policy (Halpern, 2015). The CPC2015 was primarily designed for prediction, and intuitive understanding was not a concern. Broad scope was a primary motivator of the CPC2015, with the initial problem posed by Erev et al., that multiple models have been proposed for multiple different paradigms with little overlap, leading some to want a "1-800 help line" to guide practitioners and policymakers knowledge of which model applies to which problem (Erev & Greiner, 2015). The CPC2015 organizers chose to create a new paradigm that could replicate 14 different choice anomalies in one scenario, formalizing a test of reproductive power (*3.3*) rather than broad scope as defined here.

Creating a quantitative measure of each of these criteria would not be difficult for future competitions and could be done in multiple ways. In the same way that the CPC2015 required written descriptions of models to be independently programmed, those written descriptions could be used by independent judges to provide intuitive predictions to a set of scenarios. This could be quantified differently, but the simplest suggestion would be a binary ranking of 1 (does allow intuitive predictions) or 0 (does not provide intuitive predictions). For intuitive understanding, this binary judgment could be accompanied by a basic rubric containing public policy choices like those encountered by behavioral insights teams (Halpern, 2015). Using the CPC2015 entrants as an example, variants of Prospect Theory would clearly provide intuitive predictions, as that is a major strength of Prospect Theory, while machine learning models would not necessarily provide intuitive predictions. The binary ranking of 1 would capture this theoretical advantage of Prospect Theory models over machine learning models' lack of intuitive prediction, without ruling machine learning models out of competition. Alternatively, organizers could run a parallel experiment where participants predict behavioral data from the calibration stimuli and are then given a description of a model with which to revise their predictions. A models ability to guide intuitive predictions towards accuracy would then be quantified by error reduction between the two judgments.

Broad scope could be implemented as a competition criterion a couple of different ways. Similar to the suggestion above, modeling teams could include an example of how the model could be adapted to different paradigms. This could be independently judged, with a binary 1 or 0 as above. Additionally, competition organizers could list several paradigms with models receiving a score equal to the number of paradigms they can be applied to. What is key to our ultimate comparison suggestion is an ordinal ranking of models across multiple criteria, and a model that

can be adapted to 4 different paradigms does have a theoretical advantage to a model that can only predict a single paradigm. If broad scope were a primary concern of competition organizers, an alternative would be to create stimuli in different paradigms that each model would have to predict. This instantiation would include broad scope quantitatively into MSD, possibly limiting the entry of models that do not easily adapt to different paradigms. An early competition (Erev et al., 2010) did just this, however it was presented as multiple competitions and with a few exceptions (Gonzalez & Dutt, 2011), models were only entered into one of the three possible paradigms. More complex methods of comparing models in terms of broad scope, that may not limit the entry of models that do not easily adapt, include the strong inference test (Platt, 1964) and the generalizability criterion (Busemeyer & Wang, 2000). Both of these methods involve a priori model predictions across a set of experimental conditions and would be ideal candidates for quantitatively comparing models on the criterion of broad scope.

Psychological Criteria

The psychological criteria for good models included in the previous section are: **(***2.1*) *Realistic knowledge*- model predictions should not be based on information and mental representations participants are not likely to have, (*2.2*) *Realistic capabilities*- Predictions should not rely on complex computations, and (*2.3*) *Identifiable process assumptions*- A model should rely on identifiable and testable psychological processes.

What is the benefit of developing and testing psychologically plausible models that (attempt to) capture the underlying psychological processes and the involvement of primary cognitive mechanisms such as memory, recognition, attention, etc.? Such models provide behavioral constraints on the assumptions put forward which eventually allows for quicker modifications of existing models and further developments (Johnson, Schulte-Mecklenbeck, & Willemsen, 2008). For example, Dougherty, Franco-Watkins, & Thomas (2008) mentioned that in the absence of explicit underlying processes, there are few constraints on the characteristics or features that a model can have. This then creates the problem of creating and suggesting models that are incompatible with the latent processes that they are assumed to capture.

Inspecting the outcome of any decision-making process cannot necessarily reveal how that choice was made. Additionally, identical choice profiles and outcomes can have different and distinctly identifiable psychological process profiles (e.g., Willemsen, Böckenholt, & Johnson, 2011). Identifying such basic cognitive processes (e.g., attention, memory, and reasoning) and implementing them into choice models can subsequently improve inferences about decision making (e.g., Oppenheimer & Kelso, 2015; Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011). For example, poor performance in the Iowa gambling task has been observed in a range of neuropsychological syndromes and disorders, potentially indicating a common decision-making deficit. However, with the use of cognitive models, such widespread and overly generalized deficits can be decomposed into their constituent psychological processes and provide useful and specific inferences about decision-making in clinical (or other) populations (e.g., Busemeyer, Stout, & Finn, 2003; Yechiam, Busemeyer, Stout, & Bechara, 2005). Consequently, inferences based on model parameters are dependent on the reliability and accuracy of the parameter values and whether they measure the construct that they are intended to measure. Central to this notion is the concept of parameter recoverability, that is the ability of a model (and model parameter estimation technique) to produce consistent and accurate parameter estimates (see e.g., Heathcote, A., Brown, S. D., & Wagemakers, E.-J., 2015). In addition, only when we consider psychological

and cognitive processes can we move forward in the endeavor of creating integrative and complete theories of cognition (Newell, 1973) – theories themselves cannot survive on prediction alone. A good first step to ensure that models gauge meaningful and identifiable psychological constructs or processes is to compare it against validated external measures (such as questionnaires) of the same construct that they supposedly measure (e.g., Konstantinidis, Speekenbrink, Stout, Ahn, & Shanks, 2014).

The realistic knowledge criterion naturally reflects and acknowledges the fact that any model of choice should respect and abide by the laws of ecological suitability. The main question is whether models that have been developed based on responses from laboratory participants can 1) be representative of the process or behavior they are trying to model and 2) predict behavior in more naturalistic scenarios (for relevant discussions, see Pleskac & Hertwig, 2014).

In practice, all three criteria could be ranked on a binary scale by independent judges as previously suggested. In terms of identifiable process assumptions (as we believe this is of key importance to Psychology), a competition that employed appropriate process measures such as reaction time or process tracing could provide more stringent tests of model accuracy. One additional suggestion is that modeling teams provide process predictions that are testable in future experiments.

Scientific Criteria

The final group of criteria outlined are the scientific criteria: (*3.1*) *Parsimony*- a model with the fewest meaningful parameters should be considered advantageous, (*3.2*) *Predictive power / validation* - a model should be able to predict new human behavior with accuracy, (*3.3*) *Reproductive Power* - a model should be able to reproduce common phenomena, and (*3.4*) *Testability / Falsifiability*- A model should produce predictions that could be falsified. As outlined previously, the CPC2015 was motivated by concerns of reproductive power, but used this as an exclusionary step as opposed to an evaluative criterion. The single criterion of the CPC2015 was MSD which measured predictive power, and with the split fit/prediction set up also included an aspect of parsimony, while falsifiability was not explicitly considered in the competition. Below are suggestions for how each of these criteria could be quantitatively instantiated in a competition.

Parsimony is an important scientific principle that is a desired element of models (Vandekerckhove, Matzke, & Wagenmakers, 2015). In a competition, parsimony can be incorporated into prediction metrics through a split fit/prediction set up as is the case in the CPC2015, compared to statistical measures such as AIC or BIC which penalize models with more parameters (Lewandowski & Ferrell, 2011). Alternatively, models could be rank-ordered on parsimony separately, according to the number of free parameters. The choice of how to incorporate and incentivize parsimony may depend on the exact type of competition and expected model types that may be entered. Determining a number of free parameters may be difficult as some models have assumptions that are set, but could also be interpreted as free parameters that the model builders fit absent data. Additionally, statistical models akin to machine learning may not be suitable to parameter counting in the way a typical psychological model is. A problem with some of the measures discussed thus far is that they do not penalize for functional parsimony, accounting only for parameter parsimony. To allow the most diverse types of models into a competition, a method that incorporates parsimony positively into another metric (such as MSD in the split/fit method used in the CPC2015) may be the best choice for most competitions.

Predictive power / validation is the most common (and often most important) criterion for modeling competitions. For that reason we do not need to argue one measure over another here. We anticipate some readers may initially object to our suggestion of multi-criterion competitions because they view prediction as the most (if not the only) important criterion for a model, and adding additional comparisons would only serve to weaken this one. We do not believe this is the case. As we will illustrate in the next section, in most cases prediction metrics will still be the most influential criterion for a competition winner. This results from prediction measures such as MSD providing a clear ordinal ranking of all models entered.

Reproductive power, the ability of a model to account for classic results in the literature, is a key component of a good model. As outlined earlier, reproductive power was the primary motivator of the CPC2015, with multiple models designed to account for one or two phenomena in the literature, but none being able to account for all major decision making phenomena. To address this concern the CPC2015 required models to account for 14 well known anomalies to qualify for the competition. Compiling the list of 14 choice anomalies and creating a paradigm that could replicate all of them was an endeavor that will prove useful for researchers moving forward. However the choice of using these 14 anomalies as a gate to entry as opposed to a criterion for comparison may have been a detriment to the competition. First, the 14 anomaly gate led to fewer models entered. As already stated, 53 groups registered and only 25 entered. Though there are numerous reasons for a group registering and not entering, we know of at least two groups (personal communication) that did not enter a final model because they could not account for all 14 anomalies. This is understandable, as prior to the CPC2015 no model had accounted for all 14. A second effect of making the 14 anomalies a gate as opposed to a criterion is the overabundance of BEAST variants entered into the competition. Of the 25 entrants 15 were variations of the provided baseline model. Having a difficult entry standard and providing a model that can pass that standard incentivizes teams with the goal of winning the competition to take the baseline model and attempt to improve it as opposed to entering a unique model. Not to say that the baseline model is a bad model. On the contrary, it may be a very good model, but the only conclusions that can be drawn from the competition is that it can outpredict variants of Prospect Theory, machine learning algorithms, and LIBL. If instead of using the 14 anomalies as a gate for entry, they had been used as a criterion that models were compared on, the insights from the CPC2015 could have been enhanced. For example, if a model accounted for only 13 of the 14 anomalies but contained interesting assumptions that differed from other models, those assumptions could be compared against the underlying assumptions of other models in the competition.

To use reproductive power as an evaluative criterion, a list of known effects would need to be compiled, as was done in the CPC2015, and models would be ranked in one of two ways. The first would be a simple binary ordering of models that either do or do not account for all effects. The second, used when reproductive power is a primary criterion of interest, would rank models based on the number of effects they can account for. This second way of rank ordering models also would create a situation where a true random baseline could be developed answering questions about the nature of reproductive power in the models.

Falsifiability or testability is the final scientific criterion of a good model and is one of the primary tenets of scientific advancement (Popper, 1934/1959). Such criteria need not be just theoretical, but can actually be applied in these settings. Like previous criteria, this could be a dichotomous ranking based on examples provided by competition entrants. For psychological process models, this could be relatively straightforward, such as proposing a process test such as reaction time (RT) if the model assumes RT consistent hypotheses. Falsifiability is a shortcoming

of machine learning (Russell & Norvig, 2009) which would be reflected in this criterion, though recent advances in evaluation techniques such as error terrain analysis could be argued to approximate falsifiability (Nushi, Kamar, & Horvitz, 2018).

In this section we have outlined how multiple criteria essential to what makes a good model could be incorporated into future modeling competitions. There are several advantages to using multiple criteria to evaluate models in a competition. The major advantage that we have focused on here is that it would encourage and allow more diverse models to be entered, which in turn entails more hypotheses that could be tested in a single framework. Additionally, it could improve the quality of models entered by moving away from the sole incentive of minimizing a prediction method, incentivizing the creation of models with other desirable qualities. The primary drawback of such a set up would be choosing a single winner in a manner that would be both fair and emphasize the most desirable qualities of a model (in most cases predictive accuracy). In the following section we propose a method for choosing a winner in a multi-criteria competition and argue that it is superior to current single criterion methodologies.

Choosing a Winner in a Multi-Criteria Competition

To select a winner in a multi-criteria modeling competition, we propose using a selection method from the literature on voting and computational social choice. There is a long history of research on optimal rules for selecting a winner from a series of candidates with rank dependent scoring (Goldsmith, Lang, Mattei, & Perny, 2014). Many voting rules are defined in the following way: given a voting profile P (a collection of votes, where a vote is an ordinal ranking over alternatives), each vote contributes to the score of an alternative. The global score of the alternative is then computed by summing up all these contributed ("local") scores, and finally, the alternative(s) with the highest score win(s). The most common subclass of these scoring rules is that of positional scoring rules: the local score of x with respect to vote v depends only on the rank of x in v, and the global score of x is the sum, over all votes, of its local scores. Among prominent scoring rules we find the Borda rule as well as plurality, antiplurality and k-approval (Zwicker, 2016). However, there are occasionally undesirable features of these scoring rules. Most notably, the voting rules listed can choose a winner that is not Condorcet consistent. A Condorcet method is an election method that elects the candidate that would win a majority of the vote in all of the head-to-head elections against each of the other candidates, whenever there is such a candidate. A candidate with this property is called the Condorcet winner.

In applying these rules to modeling competitions, we consider each criterion from the previous section (with ordinal rankings of the candidate models), a voting profile. In the case of modeling competitions, a Condorcet winner would beat every other model on a majority of the criteria in head to head comparisons. Because Condorcet consistency is such an important concept, we believe that it should be the primary criterion for selecting a winner in a multi-criteria competition. There are other desirable qualities of the Condorcet method though there is no need to go into them here (see Fishburn, & Gehrlein, 1977 for a detailed discussion). The primary drawback of the Condorcet methods is that there is not always a Condorcet winner. Because of this possibility, supplementary selection rules need to be agreed upon. There have been multiple suggestions for selection rules in the case where no Condorcet winner is present (e.g Peress, 2008); we believe the simplest option appropriate for the current discussion would be a Borda rule run-off, followed by a single criterion agreed upon beforehand.

To illustrate how this selection method would work practically, consider hypothetical results from two simplified competitions (Figure 2). In both competitions, the first criterion is an ordinal ranking with no ties such as MSD. The second, fourth, and fifth criteria are binary criteria with a model that satisfies the criteria ranked 1 and models that fail to satisfy the criteria ranked 2. Criterion 3 represents an ordinal ranking with ties, such as accounting for historical phenomena where a model could account for all phenomena, all but one, all but two etc.

Figure 2

Hypothetical competition rankings for two modeling competitions

Competition 1

	C1	C2	C3	C4	C5
Model 1	3	1	2	2	1
Model 2	1	1	1	1	1
Model 3	2	1	1	1	2
Model 4	4	2	1	1	2
Model 5	5	1	3	1	2

Competition 2

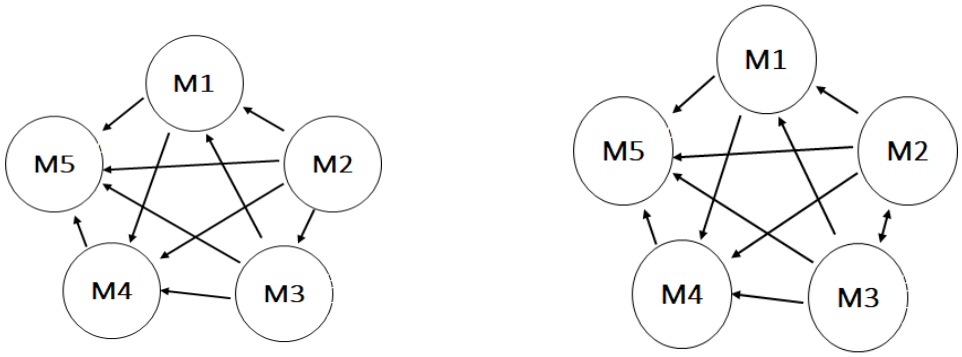
	C1	C2	C3	C4	C5
Model 1	3	1	2	2	1
Model 2	1	2	1	1	1
Model 3	2	1	1	1	1
Model 4	4	2	1	1	2
Model 5	5	1	3	1	2

Note. Figure 2 shows hypothetical rankings of 5 models across 5 criteria (C1-C5).

To first establish whether a Condorcet winner is present, all pairwise comparisons are performed with a model that is ranked above another model in a majority of criteria being superior. For example, in the first competition Model 3 is superior to Model 1 as it ranks higher than Model 1 on three out of five criteria. These pairwise comparisons can be illustrated using an edge graph (Figure 3) where a model that is superior to another has a line pointing away from it to the dominated model. A tie between models would be represented with a double headed arrow. A Condorcet winner then would have all possible lines pointing away from it. Examining Figure 3 it is clear that Model 2 is a Condorcet winner in competition 1 and would be declared the winner with no further computation.

Figure 3

Edge graphs for competitions 1 and 2



Competition 1 Competition 2

Note. Figure 3 displays an edge graph of the 5 hypothetical models from Figure 2. Directional arrows represent a model that dominates another model and double headed arrows represent a tie.

Competition 2 does not have a Condorcet winner as Models 2 and 3 are tied (each beats the other on one criterion and they are tied on the remaining three criteria). In this case a Borda run-off would be performed. In Borda rule voting, models are assigned points to their rank on each criterion with more points for higher ranks. Because Criteria 1 and 3 have more than two ranks, winners of these criteria would receive an advantage. Figure 4 shows the Borda count for each model in competition 2. In many cases a Borda run-off would determine a winner when no Condorcet winner was present. In this example however, models 2 and 3 remain tied after the Borda run off. There are two possibilities in this case; one is that organizers could agree that ties are acceptable, and two models would be declared winners; the second alternative would be using an ordinal criterion that the organizers believe to be the most important to declare a final winner. In the case of the CPC2015 and many other competitions this would be MSD.

Figure 4

Hypothetical competition rankings with a Borda run off for competition 2

Competition 2

	C1	C2	C3	C4	C5	Borda Total
Model 1	3 (3)	1 (2)	2 (2)	2 (1)	1 (2)	**10**
Model 2	1 (5)	2 (1)	1 (3)	1 (2)	1 (2)	**13**
Model 3	2 (4)	1 (2)	1 (3)	1 (2)	1 (2)	**13**
Model 4	4 (2)	2 (1)	1 (3)	1 (2)	2 (1)	**9**
Model 5	5 (1)	1 (2)	3 (1)	1 (2)	2 (1)	**7**

Note. Borda counts are in parentheses next to the original ranking.

Note that in these two fictitious examples, the model that minimized MSD more than all others is still declared the winner. The process of getting there though, opens the door for more diverse models in the competition and more methods for comparing model performance and testing auxiliary hypotheses, multiplying the potential insights that could be gained from a single competition. Additionally, the relative importance of specific criteria (i.e., prediction) could still be determined by competition organizers via binary vs. rank ordering. In the CPC2015 for example, all of the models that qualified would be ranked 1 on a reproductive power criterion, making the strictly ordinal prediction criterion more discriminating. Not only would a multi criteria competition set up improve the diversity of models entered, this more in depth model comparison procedure could clarify the best properties of the ultimate winner. In the final results of the CPC2015, 12 of the top models were statistically indistinguishable (Erev, et al., 2017, p. 389) and the winner was basically a random draw. With multiple criteria, further comparisons have the possibility of distinguishing competing models beyond their statistical tie. We present a more detailed outline of setting up and running a multi-criteria competition in supplemental online material using the CPC2015 as an example.

Discussion

Modeling competitions represent a scientific tool with great promise for advancing psychological science. The early adopters of this tool have shown that this promise can be realized in psychology. What is needed now, is a clear guide for utilizing modeling competitions in a way that meet the specific goals of psychology. We have used the CPC2015 as an example of both a laudable competition that has provided important insights into predicting human decisions and an example of how mimicking competition paradigms without considering the unique goals of psychology limits the insights that could be gained. One limitation of CPC2015 was that it was a single criterion prediction competition (i.e., MSD), rather than a multi-criteria modelling competition. Our main point in this commentary is that this limited interpretation of a modeling competition constrained the variety of models entered and subsequently the insights gained. By

starting with what is desirable in a model, we developed a list of possible quantitative criteria that could be implemented into a multi-criteria competition. Some of the benefits of this approach are 1) More diverse models would be entered into a competition, 2) The incentives of model design would be expanded beyond minimizing prediction error, 3) organizers would have flexibility to emphasize particular criteria, 4) more auxiliary hypotheses would be available for testing, and 5) the ultimate winner of the competition would be less ambiguous.

Probably the most consequential insight from CPC2015 thus far came from comparisons between BEAST and machine learning models (e.g., Plonsky, et al. 2018; Bourgin, Peterson, Reichman, Russell, & Griffiths, 2019). It is intriguing to imagine what insights could have been gained from models that did not fit one of the three model categories entered. Major decision making models such as query theory (Johnson, Häubl, & Keinan, 2007), fuzzy-trace theory (Reyna & Brainerd, 1995), and decision field theory (Busemeyer and Townsend, 1993) were not represented at all. Perhaps a larger group of process inspired models would have led to other hypotheses or combined insights like those pursued by Plonsky et al. (2018/2019). In outlining a proposed method for establishing a multi-criteria competition we have stated many of the benefits and addressed all of the possible criticisms apart from one. The one criticism of the approach we have put forth is that a competition with multiple criteria would require more work than a single criterion competition. This is absolutely true, however arranging a competition as ambitious as the CPC2015 is already a lot of work and the possible extension of insights ultimately gained from such a competition would hopefully outweigh the extra effort. The supplemental online material to this paper details one way the CPC2015 could be organized as a multi-criteria competition, and depending on the goals of the organizers the extra effort could be minimized substantially by having several of the criteria outlined here scored by the model entrants. Additionally, with a more inclusive and decisive criterion, organizers would have fewer disgruntled competitors writing commentaries and may save time on the back end!

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica: Journal of the Econometric Society*, *21*(4), 503–546. http://doi.org/10.2307/1907921

Bennett, J., & Lanning, S. (2007). The Netflix Prize. In *Proceedings of KDD cup and workshop* (pp. 3–6). New York.

Bourgin, D. D., Peterson, J. C., Reichman, D., Russell, S. J., & Griffiths, T. L. (2019) Cognitive Model Priors for Predicting Human Decisions. In *Proceedings of the36thInternational Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.

Busemeyer, J. R., & Diedrich, A. (2009). *Cognitive Modeling*. California: Sage Publications Inc.

Busemeyer, J. R., Stout, J. C., & Finn, P. (2003). Using computational models to help explain decision making processes of substance abusers. In D. Bargh (Ed.), *Cognitive and affective neuroscience of psychopathology*. Oxford, England: Oxford University Press.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459.

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

Drivendata.org

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review*, *115*(1), 199–211. https://doi.org/10.1037/0033-295X.115.1.199

Erev, I., Ert, E., Plonsky, O., Cohen, D., & Cohen, O. (2017). From anomalies to forecasts: Toward a descriptive model of decisions under risk, under ambiguity, and from experience. *Psychological Review*, *124*(4), 369–409. https://doi.org/10.1037/rev0000062

Erev, I., Ert, E., & Roth, A. (2010). A Choice Prediction Competition for Market Entry Games: An Introduction. *Games*, *1*(2), 117–136. https://doi.org/10.3390/g1020117

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., … Lebiere, C. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15–47. https://doi.org/10.1002/bdm.683

Erev, I., Greiner, B. (2015). The 1-800 Critique: Couterexamples, and the Future of Behavioral Economics. In G. R. Fréchette, and A. Schotter, A (Ed.) *Handbook of Experimental Economic Methodology*, edited by Oxford: Oxford University Press, 151-165.

Fishburn, P. C. & Gehrlein, W. V. (1977). *An Analysis of Voting Procedures with Nonranked Voting*. Behavioral Science 22:178-85.

Goldbloom, A. (2010). Data Prediction Competitions -- Far More than Just a Bit of Fun. In *2010 IEEE International Conference on Data Mining Workshops* (pp. 1385–1386). IEEE. https://doi.org/10.1109/ICDMW.2010.56

Goldsmith, J., Lang, J., Mattei, N., & Perny, P. (2014). Voting with rank dependent scoring rules. In *Proceedings of the National Conference on Artificial Intelligence* (Vol. 1, pp. 698–704).

Gonzalez, C. (2017). Decision making: A cognitive science perspective. In Chipman, S. (Ed), The Oxford Handbook of Cognitive Science 249-263. New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780199842193.013.6

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating decisions from experience in sampling and repeated choice paradigms. *Psychological Review, 118*(4), 523-551. DOI:10.1037/a0024558

Gonzalez, C., Dutt, V., & Lejarraga, T. (2011). A loser can be a winner: Comparisons of two instance-based learning models in a market entry competition. Games, 2, 136–162.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4), 591-635. DOI:10.1016/S0364-0213(03)00031-4

González-Vallejo, C., Harman, J. L., Mullet, E. & Muñoz-Sastre, M. T. (2012). An examination of the proportional difference model to describe and predict health decisions. *Organizational Behavior and Human Decision Processes.* No. 118, p. 82-97.

Halpern, D. (2015). *Inside the Nudge Unit: How Small Changes Can Make a Big Difference*. London: Penguin Random House.

Heathcote, A., Brown, S. D., Wagemakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In Forstmann, B. U., & Wagenmakers, E.-J. (Eds.), An introduction to model-based cognitive neuroscience (pp. 25-48). New York, NY: Springer, New York.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539. DOI:10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517-523. DOI:10.1016/j.tics.2009.09.004

Hertwig, R. (2015). Decisions from experience. The Wiley Blackwell handbook of judgment and decision making, 1, 240-267.

Hunt, N. (2014) Quantifying the Value of Better Recommendations. *ACM Association for Computing Machinery Recommender Systems Conference* (RecSys).

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33, 461 – 474.

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review*, *115*(1), 263–272. https://doi.org/10.1037/0033-295X.115.1.263

Kaggel.com

Katsikopoulos, K. V. **(**2014). Bounded Rationality: The Two Cultures." *Journal of Economic Methodology* 21: 361-74.

Katsikopoulos, K. V. (2020). The merits of transparent models. *Behavioral Operational Research*. P.261-275.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, *47*(2), 263–292. http://doi.org/10.2307/1914185

Konstantinidis, E., Speekenbrink, M., Stout, J. C., Ahn, W.-Y., & Shanks, D. R. (2014). To simulate or not? Comment on Steingroever, Wetzels, and Wagenmakers (2014). *Decision*, *1*(3), 184–191. https://doi.org/10.1037/dec0000016

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.),The essentialtension(pp. 320339). Chicago: University of Chicago Press.

Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lebiere, C., Gonzalez, C., & Warwick, W. (2010). Cognitive architectures, model comparisons, and AGI. Journal of Artificial Intelligence, 2(2), 1-19.

Lewandowski, S. and Farrell, S. (2011). *Computational Modeling in Cognition: principals and practice*. SAGE Publications.

Meir, R., Lev, O., & Rosenschein, J. S. (2014). A local-dominance theory of voting equilibria. In *Proceedings of the 15th ACM Conference on Economics and Computation*, 313–330.

netflixprize.com

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 283–308). New York, NY: Academic Press.

Nushi, B., Kamar, E., & Horvitz, E. (2018). Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. *ArXiv*, *abs/1809.07424*.

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*(1), 277–294. https://doi.org/10.1146/annurev-psych-010814-015148

Peress, M. Selecting the Condorcet Winner: single-stage versus multi-stage voting rules. *Public Choice* 137**,** 207–220 (2008). https://doi.org/10.1007/s11127-008-9321-y

Platt, J. R. (1964). Strong Inference. *Science, 146(3642), 347-353.* Doi: 10.1126/science.146.3642.347

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000–2019. https://doi.org/10.1037/xge0000013

Plonsky, O., Apel, R., Erev, I., Ert, E., and Tennenholtz,M. (2018). When and how can social scientists add value to data scientists? A choice prediction competition for human decision making. Unpublished Manuscript.

Plonsky, O., Apel, R., Ert, E., Tennenholtz, M., Bourgin, D., Peterson, J. C., Reichman, D., Griffiths, T. L., Russell, S. J., and Carter, E. C. (2019) Predicting human decisions with behavioral theories and machine learning. arXiv preprintarXiv:1904.06866, 2019.

Plonsky, O. Erev, I. Hazan, T. & Tennenholtz, M. (2017). Psychological forest: Predicting human behavior. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI, 656–662. AAAI Press.

Popper, (1959). *The logic of scientific discovery*. New York: Basic Books. (Original work published 1934)

Reyna & Brainerd (1995). *Fuzzy-trace theory: An interim synthesis.* Learning and Individual Differences, 7 (1995), pp. 1-75,

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*(2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach.* 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Sajda, P., Gerson, A., Muller, K., Blankertz, B., & Parra, L., (2003) A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 184-185, June 2003.

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making*, *6*(8), 733–739.

Spiliopoulos, L., & Ortmann, A. (2014). Model comparisons using tournaments: Likes, "dislikes," and challenges. *Psychological Methods*, *19*(2), 230–250. https://doi.org/10.1037/a0034249

Triguero, I., del Río, S., López, V., Bacardit, J., Benítez, J. M., & Herrera, F. (2015). ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem. *Knowledge-Based Systems*, *87*, 69–79. https://doi.org/10.1016/j.knosys.2015.05.027

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*(2), 269–283.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *Oxford library of psychology. The Oxford handbook of computational and mathematical psychology* (p. 300–319). Oxford University Press.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. New York, NY: Cambridge University Press.

Weber E. U. & Johnson, E. J. (2012) Mindful Judgment and Decision Making. *Annual Review of Psychology.* 60,53-85

Weigend, A. S., & Gershenfeld, N. A. (1993) Results of the time series prediction competition at the Santa Fe Institute. *IEEE International Conference on Neural Networks*.

Willemsen, M., Böckenholt, U., & Johnson, E. J. (2011). Choice by value encoding and value construction: Processes of loss aversion. *Journal of Experimental Psychology: General*, *140*(3), 303–324. https://doi.org/10.1037/a0023493

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science*, *16*(12), 973–978. https://doi.org/10.1111/j.1467-9280.2005.01646.x

Zwicker, W. S., (2016). Introduction to the theory of voting. In F. Brandt, V. Conitzer, U. Endriss, J. Lang, & A. D. Procaccia (Eds). *Handbook of Computational Social Choice*, (p. 23-56). Cambridge University Press.