# The Wisdom of Model Crowds

**Preprint** · May 2021

DOI: 10.31234/osf.io/8ygvp



**3 authors**, including:

Pantelis P. Analytis
University of Southern Denmark
**25** PUBLICATIONS **432** CITATIONS

SEE PROFILE

# The Wisdom of Model Crowds

Lisheng He
Shanghai University and Shanghai International Studies University

Pantelis P. Analytis
University of Southern Denmark

Sudeep Bhatia
University of Pennsylvania

May 26th, 2021

Correspondence should be addressed to Lisheng He, SILC Business School, Shanghai University, Shanghai China. Email: felix8.he@gmail.com.

**Acknowledgement**

We would like to thank Yuval Rottenstreich, two anonymous reviewers and the attendees at MathPsych (2019), SPUDM (2019), Psychonomics (2019) and SJDM (2019) meetings for their insightful comments on earlier drafts, and the authors who kindly shared their data with us.

Funding for Lisheng He was received from the Shanghai Pujiang Program (2020PJC102) and the Fundamental Research Funds for the Central Universities (2020114083). Funding for Sudeep Bhatia was received from the National Science Foundation grant SES-1847794 and the Alfred P. Sloan Foundation.

**Abstract**

A wide body of empirical research has revealed the descriptive shortcomings of expected value and expected utility models of risky decision making. In response, numerous models have been advanced to predict and explain people's choices between gambles. Although some of these models have had a great impact in the behavioral, social and management sciences, there is little consensus about which model offers the best account of choice behavior. In this paper, we conduct a large-scale comparison of 58 prominent models of risky choice, using 19 existing behavioral datasets involving more than 800 participants. This allows us to comprehensively evaluate models in terms of individual-level predictive performance across a range of different choice settings. We also identify the psychological mechanisms that lead to superior predictive performance and the properties of choice stimuli that favor certain types of models over others. Second, drawing on research on the wisdom of crowds, we argue that each of the existing models can be seen as an expert that provides unique forecasts in choice predictions. Consistent with this claim, we find that crowds of risky choice models perform better than individual models and thus provide a performance bound for assessing the historical accumulation of knowledge in our field. Our results suggest that each model captures unique aspects of the decision process, and that existing risky choice models offer complementary rather than competing accounts of behavior. We discuss the implications of our results on theories of risky decision making and the quantitative modeling of choice behavior.

Keywords: Decision making; risky choice; crowd wisdom; model ensembles; choice prediction

## Introduction

Risk plays a key role in everyday choice, with managerial, financial, consumer, and health decision making often involving the evaluation of probabilistic outcomes, and the optimization of value in the face of uncertainty. Unsurprisingly, understanding how people make risky choices is one of the most important research topics in the behavioral sciences, and a central focus of fields such as managerial decision making, behavioral decision research, judgment and decision making, and behavioral economics. Dating back to the correspondence between Blaise Pascal and Pierre de Fermat, some have argued that people *should* always maximize expected value in risky choice. However, early thought experiments, such as the St. Petersburg paradox proposed by Nicolaus and Daniel Bernoulli (1738), have challenged this view by speculating what people *would* do, therefore putting alternative descriptive theories of decision under risk into perspective. Following these challenges, expected utility theory (EUT), pioneered by Daniel Bernoulli (1738) and axiomatized by von Neumann and Morgenstern (1944), has provided an influential approach to thinking about both normative and descriptive aspects of risky choice.

Of course, research on risky choice behavior did not end with EUT. Rather, the question of what people do when confronted with options that offer potentially probabilistic outcomes has fueled a transgenerational, interdisciplinary research program, with tremendous impact both within academia and in applied settings. The first behavioral experiments designed to answer this question focused on specific deviations from EUT (Allais 1953, Edwards 1954). Soon several discrepancies had been uncovered, and the accumulated empirical evidence gave rise to a wave of fully-fledged behavioral models (each associated with different psychological mechanisms) that could be directly contrasted to EUT in terms of descriptive adequacy (e.g. Kahneman and

Tversky 1979, Busemeyer and Townsend 1993, Birnbaum 2008). The rate at which new models have been advanced has only accelerated over the years---at the time of writing this article several dozens of behavioral models of risky choice had been proposed (Starmer 2000, He et al. 2020).

Judging by the volume of models available to explain existing data, the study of people's risk-taking behavior should be one of the most mature fields in the social and behavioral sciences. What is the current state-of-the-art in terms of describing people's behavior and predicting their choices? How much progress have we collectively achieved across disciplines and what are the psychological mechanisms that are necessary to get good predictions? Surprisingly, it is hard to find answers to these questions. More often than not, different models are seen as competitors, where the success of a model directly discredits rival theoretical accounts. Moreover, it remains hard to assess the relative importance of different psychological mechanisms and the overall output of the collective scientific endeavor, as the study of risky choice is rather fragmented even within disciplines, and even more so across disciplines.

There are three main roadblocks hindering progress and synthesis across disciplines. First, new theoretical papers typically compare the advanced model against a handful of main competitors; as a result, it is hard to judge how a model fares against the overall state-of-the-art in predicting and describing people's choices. Although this is a reasonable approach given the large number of potential competitors, it can lead to a splintered view of the literature and important ideas being forgotten. Second, different studies use very different datasets to evaluate the performance of models. Model performance largely depends on the selection of stimuli included in different experiments (see Erev et al. 2017 for a similar critique) and consequently the predictive ability of models varies across studies, making comparisons between different

theoretical accounts particularly complicated. Finally, among the existing empirical studies only a modest subset have generated enough data to allow for the estimation of model parameters of individuals, despite the fact that model parameters correspond to psychological factors, such as subjective perception, attention and emotion, which could be highly idiosyncratic across people (Edwards 1955, Bordalo et al. 2012, Loewenstein et al. 2015). In the absence of such individual-level tests, our understanding of the descriptive power of many existing models is incomplete.

What is needed is a trans-disciplinary analysis that comprehensively integrates the rich set of theoretical insights identified by prior researchers, and uses these insights to identify the state-of-the-art in modeling individual-level risky choice, quantify the progress made over the past several decades, understand how key psychological properties of these models relate to model performance for different datasets, and develop novel ideas for improving the predictive and explanatory scope of risky decision making research. In this paper we hope to present such an analysis. First, we build a collection of 58 risky choice models from numerous papers published in disciplines such as management, economics, and psychology. Importantly, we instantiate these models in code, thereby formalizing their functional forms and rigorously specifying their implementation details. To the best of our knowledge, this is the most extensive set of risky choice models compiled and implemented so far. Second, we build a collection of risky choice datasets, again drawn from different papers. Our collection includes both datasets with mixed gambles and with only positive gambles (i.e. gains), and datasets with numerous different types of choice problems (including randomly generated and experimenter-curated choice problems, one and two non-zero branches choice problems), allowing for a much more comprehensive evaluation of different models. Additionally, each of our datasets has a large number of responses on the individual level, facilitating individual-level model fits and tests.

Overall, these datasets involve 825 individuals making 76,910 risky choices in total. Again, to the best of our knowledge, this is one of the largest risky choice datasets compiled so far. The large panel of models and the vast test-bed of datasets allow for an unprecedentedly complete evaluation of different individual models. In fact, we are the first to quantitatively fit many of the models in our dataset, and our tests outperform the size of the datasets and model sets used in prior work by an order of magnitude.

The rich collection of models and choice stimuli that we have collected also allows us to better understand the properties of the models and choice problems that drive our results. We attempt this analysis by partitioning our set of models based on the assumed psychological mechanisms (e.g. probability weighting, regret, attention etc.) and by partitioning our stimuli based on the correlations between the underlying probabilities and payoffs as well as the expected value (EV) difference between options. We are subsequently able to test which mechanisms lead to superior model performance and how this varies based on the underlying stimuli structure offered to participants.

Seeing models as competing against each other is a limitation in itself, and may not do full justice to the historical accumulation of knowledge in risky choice research. A more productive approach may be to consider models as complementary and thereafter to exploit the collective wisdom accumulated in different research papers across different disciplines. Thus, in this paper, we integrate research on risky choice and research on the wisdom of crowds (Galton 1907, Surowiecki 2004) to develop and test *model crowds*, where each model is seen as an expert whose judgments can be aggregated with those of other models to better describe choice behavior. Hitherto, the principle of crowd wisdom has been used to aggregate the opinions of different people in estimation and categorization tasks. Averaged opinions typically lead to more

reliable estimates, and in many cases outperform the predictions of the best-performing individual. In a similar vein, model crowds could leverage insights of various risky choice models and predict people's behavior better than any individual model. Moving from individuals to models is a natural step. In fact, there are often towering intellectual figures standing behind the models, and models can in many ways be seen as the (mathematically specified) decision rules that would be used by these experts to predict individual choice.

Model crowds hold great promise for improving our ability to predict people's behavior. In the field of machine learning, model aggregation has proven valuable for improving prediction in regression and classification tasks by efficiently leveraging small amounts of data and reducing sensitivity to specific samples (and thus reducing variance, Breiman 1998; Polikar, 2006). Thus, it comes as no surprise, that ensemble models which aggregate the predictions of several distinct models, are often proclaimed the winners of machine learning competitions (Bell and Koren 2007, Niculescu-Mizil et al. 2009). Closer to home, ensemble models have shown great promise in a series of prediction competitions featuring models that were developed and tuned by research teams using training data from large behavioral experiments with the goal to predict the proportion of people choosing a risky option over another in a hold-out dataset (Erev et al. 2010, Erev et al. 2017) and have been leveraged by cognitive modelers to uncover people's cognitive processes (Singmann et al. 2018). What's more, in risky choice and other choice processes more broadly, it is reasonable to assume that individuals' decision strategies may be governed by a number of factors that are not present in any single decision model. Thus, crowds of individual models relying on different theoretical assumptions may capture these factors and thus predict individual level behavior better than any single model does (Payne et al. 1988, Scheibehenne et al. 2013).

Overall, model crowds combine the insights of numerous existing models to predict and describe choice behavior, and thus provide a measure of the progress we have collectively achieved across disciplines. We can also quantify the contribution of each individual model in a model crowd, which can be used to identify the idiosyncratic predictive value of the model (when taking into account the predictions of other models in the crowd). When crowd models are evaluated over entire historical periods they can be used to quantify the growth of knowledge over time, becoming a powerful tool to study the history of risky choice research. Last but not least, by calculating the average weights of models that rely on a specific psychological mechanism or by removing all models using a specific mechanism, models crowds can provide a measure of the relative importance of different psychological mechanisms in improving our predictive ability. Below we test our 58 risky choice models, along with various model crowds generated from these models, as well as their assumed psychological mechanisms, to obtain a comprehensive understanding of the descriptive power of behavioral theories of risky choice.

## Methods

### Models

We collected the long list of risky choice models using a multistage process. We first searched Google Scholar using various keywords (e.g. *risky choice model*, *risky decision model*), and looked for models in regular review articles published in the Annual Review of Psychology and the Journal of Economic Literature (Edwards 1954, 1961, Becker and McClintock 1967, Rapoport and Wallstern 1972, Slovic et al. 1977, Einhorn and Hogarth 1981, Pitz and Sachs 1984, Payne et al. 1992, Simonson et al. 2001, Starmer 2000, Hastie 2001, Weber and Johnson 2009, Oppenheimer and Kelso 2015). Then, using citation chaining we found additional models presented in papers citing our list of models. We then circulated these models to our colleagues

using the Society for Judgment and Decision Making email listserv, who helped us identify additional models not present in our list. Finally, we manually searched through prominent journals in management, psychology and economics, such as Management Science, Psychological Review, and American Economic Review, for recently published models that may not have been on our list.

Overall, our focus was on mathematically or algorithmically specified models of description-based risky choice with precise functional forms that could be fit to choice data. Thus, we excluded models of decision making under ambiguity (e.g. Camerer and Weber 1992), models of experience-based risky decision making (e.g. Hertwig and Erev 2009, Gilboa and Schmeidler 1995), models of reference dependence (e.g. Kőszegi and Rabin 2006), qualitative models (e.g. Loewenstein et al. 2001), purely axiomatic models without restrictions on functional forms (e.g. Machina 1982), models of risk perception (e.g. Pollatsek and Tversky 1970), and models that did not have analytically specified likelihood functions and needed to be simulated to make predictions (e.g. Erev et al. 2017).

Despite these restrictions, we were able to collect 58 distinct models. Each of these models makes implicit or explicit assumptions about the psychological mechanisms at play in risky choice, and our large collection of models gives us an unprecedented opportunity to analyze the role of these mechanisms in model performance. After consulting the original papers of the models and identifying the mechanisms that their authors evoked when presenting the models, we categorized models as involving one or more of nine mechanisms: (1) payoff transformation, (2) probability transformation, (3) attention, (4) sampling, (5) regret, (6) disappointment, (7) ranking, (8) threshold and (9) dispersion (see Figure 1). Models with the first and second mechanisms transform payoffs into subjective values (e.g. Bernoulli 1738) or

probabilities into subjective probabilities (e.g. Edwards 1954) using non-linear functions, and use these transformed payoffs or probabilities to evaluate the gambles. Models with attention (e.g. Busemeyer and Townsend 1993, Birnbaum 2008) assume that decision makers selectively focus on some payoffs, probabilities, or states of the world, whereas models with sampling (e.g. Lieder et al. 2018) assume that decision makers simulate or retrieve from memory the outcomes that are used to evaluate the gambles. Models that allow for regret (e.g. Bell 1982, Loomes and Sugden 1982) typically compare the payoffs of a gamble against the payoffs of other gambles, whereas models that allow for disappointment (e.g. Bell 1985, Loomes and Sugden1985) typically compare the payoffs of a gamble against the payoffs of the same gamble. Models that use ranking (e.g. Thorgate 1980, Birnbaum 1997), order the payoffs or probabilities involved and make decisions based on the ranks of these payoffs or probabilities. Models that use thresholds (e.g. Fishburn 1977, Diecidue and van de Ven 2008) typically use discrete cutoffs for payoffs or probabilities to evaluate gambles. Models with the dispersion mechanism (e.g. Markowitz 1952, Weber 2004) compute some measure of variability for gambles, and typically penalize models with high variance payoffs. Of course a given model can allow for multiple mechanisms at the same time, such as transformations of both payoffs and probabilities (e.g. prospect theory, see Kahneman and Tversky, 1979), decision making under the influence of both regret and disappointment (e.g. Mellers et al. 1999), or heuristic choice with a sequence of transformation and threshold operations (e.g. Leland 1994). These mechanisms are non-exclusive, and a certain model may make use of more than one of them.

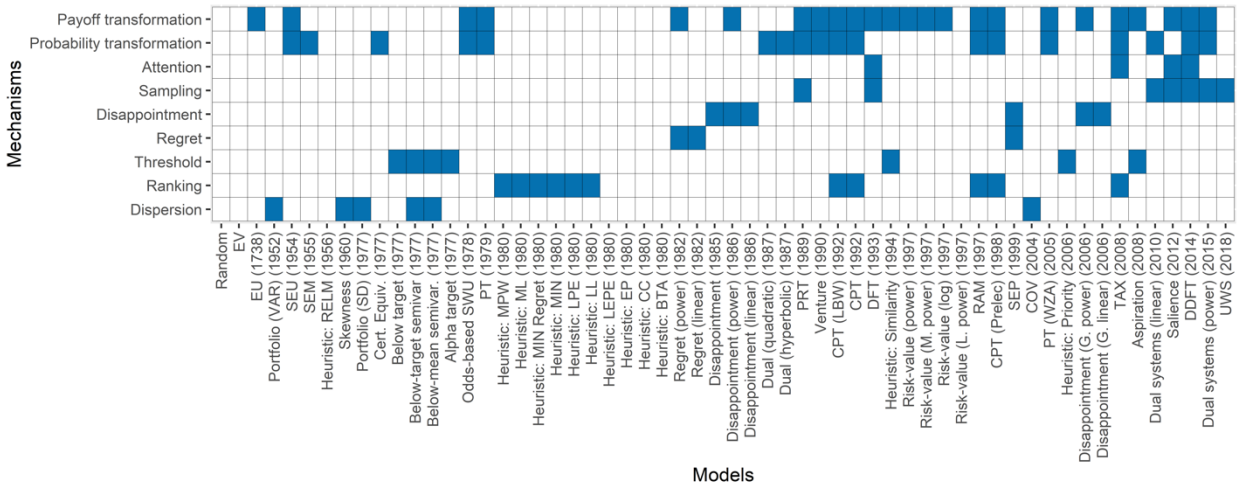


Figure 1. Psychological mechanisms in risky decision models. Blue shades mean that the model involves the psychological mechanism. Details and full names of the models can be found in the Appendix.

In order to fit stochastic choice data, we applied the Logit choice rule to models that generate utilities or choice propensities on a cardinal scale. For models that generate choice propensities on an ordinal scale (e.g. heuristics), for which the Logit rule was not applicable, a trembling-hand (i.e. constant-error) choice rule was applied to accommodate choice stochasticity. The two stochastic specifications are not as different as they may appear. Indeed, if we allow the Logit choice rule to take an ordinal preference order as the input, it reduces to the trembling-hand choice rule (by yielding a probability of choosing the preferred option, and a complementary probability of making an error fixed across items). Additional details regarding the models and their implementation are provided in the Appendix.

**Datasets**

We evaluated the models with a wide range of datasets from experimental studies. First, we downloaded datasets that have been made available online (either on personal websites or public repositories) from recently published papers with risky choice experiments. We also sent an email to the Society for Judgment and Decision Making listserv, requesting relevant datasets. All datasets were further screened to meet the following criteria: 1. Individual-level choice data

with at least 50 choice problems for each participant (as described in the request email); 2. Binary choice between monetary gambles (with explicit descriptions for both probabilities and payoffs); 3. At most two possible monetary outcomes for each gamble.

With the above measures taken, we obtained a total of 19 datasets (see Table 1). Twelve of these datasets involved gambles purely in the gain domain, including one originally presented in Rieskamp (2008), two in Fiedler and Glöckner (2012), eight in Stewart et al. (2015) and one in Stewart et al. (2016). Rieskamp's (2008) dataset involved 30 participants making 60 binary risky choices each. Stewart et al.'s (2015) datasets involved a total of 208 participants, each of whom made either 120 or 150 binary risky choices. Stewart et al. (2016) involved 48 participants and each participant made 71 choices. The other seven datasets involved gambles with both gains and losses (i.e. mixed gambles), including one dataset collected by Erev et al. (2017), one by Pachur et al. (2017) and five by Pachur et al. (2018). Erev et al.'s (2017) dataset involved 60 participants making 57 binary choices each. Pachur et al.'s (2017) dataset involved 122 participants making 105 binary choices each. Pachur et al.'s (2018) datasets involved 300 participants making either 91 or 51 binary choices each. Overall, the full array of datasets involved 343 participants making 38,180 choices in the gain domain and 482 participants making 38,730 choices in the mixed domain. Note that four models (the relative risk-value models, Dyer and Jia 1997) were designed exclusively for risky choice in the gain domain and thus were excluded for the analysis of mixed gambles (which involved losses). Thus, there were 58 models for gains and 54 models for mixed gambles. As such, the results from the two types of datasets are presented separately below. As reported in the original papers, participants in these experiments were incentivized based on their choices in the tasks.

Table 1. Summary of the datasets for model evaluation

| Source | Abbreviation | Type | Gamble design | # None-zero | # Datasets | # Participants | # Trials |
|---|---|---|---|---|---|---|---|
| Rieskamp (2008) | Rieskamp08 | Gains | Random | 2 | 1 | 30 | 60 |
| Fiedler and Glöckner (2012) | FG12 | Gains | Random/Manual | 2 | 2 | 57 | 50 |
| Stewart et al. (2015) | SRH15 | Gains | Manual | 1 | 8 | 208 | 120 or 150 |
| Stewart et al. (2016) | SHM16 | Gains | Manual | 1 | 1 | 48 | 71 |
| Erev et al. (2017) | EEPCC17 | Mixed | Random | 2 | 1 | 60 | 57 |
| Pachur et al. (2017) | PMH17 | Mixed | Random/Manual | 2 | 1 | 122 | 105 |
| Pachur et al. (2018) | PSMH18 | Mixed | Random/Manual | 2 | 5 | 300 | 91 or 51 |
| Total | | - | - | - | 19 | 825 | - |

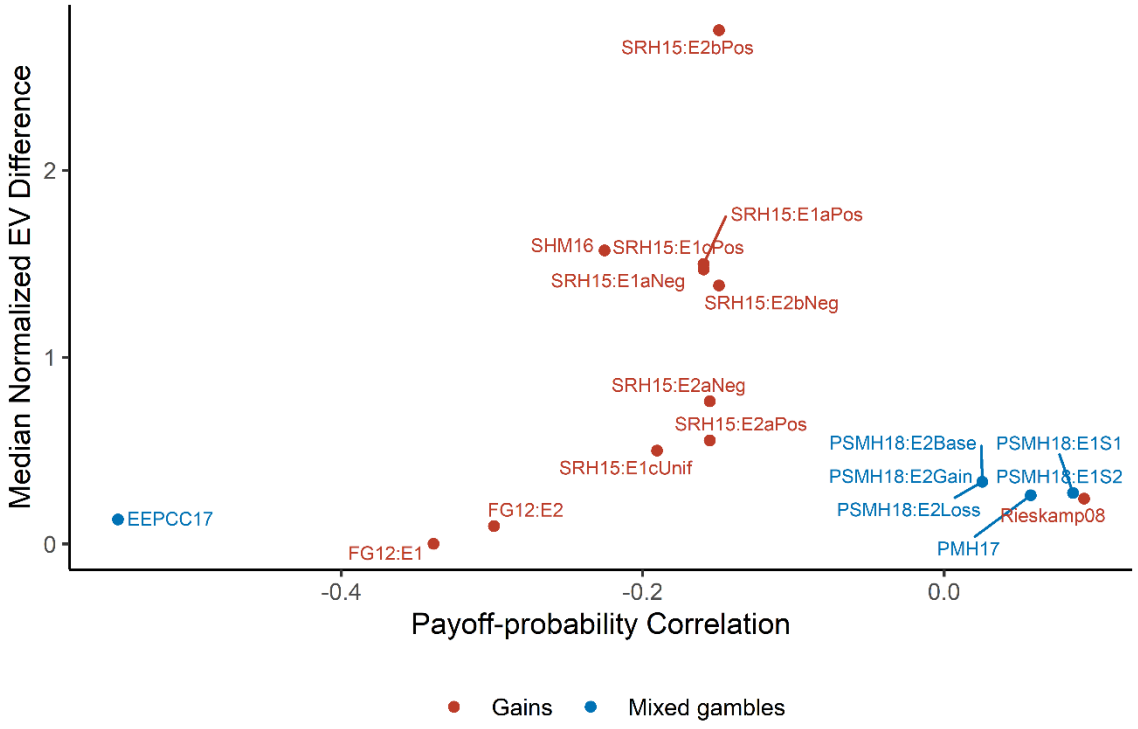


Figure 2. Two-dimensional display of the datasets based on the payoff-probability correlation and the median value of the normalized EV differences. The labels of the datasets can be found in the Abbreviation column of Table 1. (see the color figure online)

The datasets compiled in this paper involve a wide range of gamble designs. Some datasets have generated gambles by systematically crossing payoffs with probabilities and exhausting all possible combinations of payoffs and probabilities (e.g. Stewart et al. 2015, 2016) while others have randomly selected gambles from a reasonable stimulus space (Erev et al. 2017). Some designs have featured items that people commonly encounter in real-world settings (Rieskamp 2008). Yet others have followed a hybrid approach that combines manually crafted gambles with randomly generated gambles (e.g. Pachur et al. 2018). These designs can also be understood in terms of two key quantitative properties: (1) the normalized expected-value (EV) difference between options and (2) the correlation between payoffs and its associated probabilities. The normalized EV difference is a choice-level property. For each binary choice between *X* and *Y*, the normalized EV difference is defined as $\frac{|EV_X - EV_Y|}{\min\{EV_X, EV_Y\}}$. The correlation between payoffs and probabilities is an experimental dataset-level property. It is defined as the Pearson's correlation between all involved payoffs and their associated probabilities in the experiment. Note that some researchers have intentionally controlled the normalized EV differences in the stimuli sample to allow for data-efficient model selection (e.g. Rieskamp 2008).

In Figure 2 we plot each dataset in our analysis in terms of the median normalized EV difference of its component choice problems, as well as the correlation between payoffs and probabilities across all its choice problems. We see here that most of our datasets have a negative correlation between payoffs and probabilities, corresponding to a more ecologically valid design (Pleskac and Hertwig 2014). There is a large amount of variance in the median normalized EV difference across datasets, and only a few datasets keep this difference fixed at zero or very close

to zero. There do not appear to be systematic differences between gains and mixed gambles on these two dimensions.

**Cross validation**

The 58 models in our collection have varying numbers of parameters and different assumptions leading to different degrees of flexibility. To control for flexibility, we used 10-fold cross-validation and evaluated the models' out-of-sample predictive performance. All the analyses were conducted at individual level. Each individual participant's choice data were divided into ten subsets. In each iteration, nine subsets (i.e. 90% of the choice data) served as the training set to train the models and estimate their free parameters and the remaining subset served as the test set. The training-testing procedure was repeated ten times for each participant, with each of the ten subsets serving as the test set once. Parameters were estimated by means of maximum likelihood (Pitt et al. 2003). To ensure that global maximum was reached, we repeated the SIMPLEX algorithm 500 times in the MATLAB *fminsearch* function and selected the maximum likelihood estimation. For a given model *m*, the estimated parameters in the training set were used to make predictions in the test set. For each choice problem *i*, the out-of-sample prediction using these parameters is denoted as $\hat{y}_{m,i}$, which is the predicted probability that the first of the two options on the choice problem is chosen. Because each trial served in the test set exactly once in the 10-fold cross validation, we obtained an out-of-sample prediction for every choice problem in each participant's choice data.

We evaluated models' out-of-sample predictive performance with three different loss functions. The first one is the binary prediction error. For a given model *m*, the individual-level prediction error is defined as $PE_m = \frac{1}{N}\sum_{i=1}^{N} I\left(\left|\hat{y}_{m,i} - y_i\right| > 0.5\right)$, where $y_i$ is the observed choice (1 if the first option is chosen on problem *i*, 0 otherwise) and *N* is the number of choice

problems. $\mathrm{I}(\cdot)$ is the indicator function that returns 1 if the argument is true, and 0 otherwise. Note that in the rare cases where the prediction $\hat{y}_{m,i}$ was exactly 0.5, the indicator function $\mathrm{I}(\cdot)$ was replaced with a prediction error of 0.5. The others were two probabilistic loss functions: Log-loss and Brier Score. The log-loss is defined as $LL_m = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_{m,i}) + (1 - y_i)\log(1 - \hat{y}_{m,i})\right]$. The Brier score is defined as $BS_m = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_{m,i} - y_i\right)^2$. The smaller the errors according to these loss functions, the better the model performance. The probabilistic loss functions take into account the strength of preferences and are thus more sensitive to the models' quantitative predictions than the prediction error, which by contrast only encodes the direction of preference.

**Model crowds**

In addition to individual models, we built and tested five model crowds, inspired by research on the wisdom of crowds. Our model crowds took the individual models' (trained) out-of-sample predictions on the test set as given and then made novel predictions by combining the individual model predictions using some model weighting scheme. Importantly, in designing such model crowds, we assigned weights to the models independently of the test set, which remained fully out-of-sample. As with the individual models, model crowds were evaluated based on their out-of-sample predictions with the same set of loss functions described above.

The first model crowd used in our analysis was a naïve crowd that unconditionally averages out the predictions of all models for each choice problem in the test set. Specifically, for each choice problem *i* in the test set, the naïve crowd's predicted choice probability is the unweighted average of all individual models' choice probabilities: $\hat{y}_{nc,i} = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_{m,i}$, where *M* is the total number of individual models. Intuitively, the naïve crowd sees each individual model as

being an equally valid predictor and thus aggregates the individual models without weights (as with, for example, the equal weights heuristic decision rule, Dawes et al. 1989). Despite its simplicity, this model has been shown to perform quite well in forecasting opinion aggregation contexts, largely due to the robustness (low variability) of its predictions (Hogarth 1978, Clemens 1989, Armstrong 2001, Analytis et al. 2018).

A second model crowd was the weighted crowd. This model used differences in model performances at the training stage to inform model weights, so that that better-performing models at the training stage were given higher weights in the crowd. We used Akaike weights for this purpose (Akaike 1973, Wagenmakers and Farrell 2004). The Akaike weight for a model is proportional to the model's maximum likelihood in the data it is fit on (in our case, the training data), but also includes a penalty for model complexity in terms of the number of free parameters. Accordingly, for each choice problem *i* in the test set, the weighted crowds' predicted choice probability is: $\hat{y}_{wc,i} = \sum_{m=1}^{M} w_m^{Akaike} \hat{y}_{m,i}$, where $w_m^{Akaike}$ is model *m*'s Akaike weight, with $\sum_{m=1}^{M} w_m^{Akaike} = 1$. Akaike weight is defined as $w_m^{Akaike} = \frac{\exp(-\frac{1}{2} AIC_m)}{\sum_{k=1}^{M} \exp(-\frac{1}{2} AIC_k)}$, where $AIC_m = -2 \log L_m + 2V_m$ is the Akaike Information Criterion for the training set ($\log L_m$ is the maximum log likelihood of the training data and $V_m$ is the number of free parameters in *m*). The weighted crowd model can be seen as aggregating individual model predictions in a way that places more emphasis on the predictions of models that perform well on the training data, and thus models whose predictions are more likely to be correct in the test data (as with, for example, the weighted additive decision rule, e.g. Keeney and Raiffa 1993). Similar models in the wisdom-of-crowds literature weigh the predictions of individuals based on their accuracy in prior

forecasts, their self-reported confidence, or some other measure of individual-level performance, and for this reason often outperform the naïve crowd (Einhorn et al. 1977, Armstrong 2001, Bahrami et al. 2010). The weighted crowd can also be seen as an alternative implementation of Bayesian model averaging that penalizes model complexity by means of the number of free parameters (Hoeting et al. 1999).

Our third and fourth model crowds were select crowds (Mannes et al. 2014, Goldstein et al. 2014). As with the weighted crowd, the select crowds utilize differential model performance in the training set to determine model weights for predictions for the test set. They identify a particular number of best-performing models in the training set and assign an equal weight to all selected models. Consistent with several recent applications of select crowds we varied the crowd size (e.g. Luan et al. 2012, Mannes et al. 2014, Goldstein et al. 2014, Analytis et al. 2018, Galesic et al. 2018). Specifically, we selected either the top-five or top-ten best performing models in the training set for each training-testing iteration and obtained the select crowd's predictions by unconditionally averaging the predictions of the selected models. These models are referred to as select-5 and select-10 crowds respectively.

Aggregating the opinions of five or near to five models or experts has been shown to lead to good results across settings (Makridakis and Winkler 1983, Ashton and Asthon 1985). The prediction improvement tends to diminish as additional models or experts are added to the select crowd, depending on the quality of information about past judge performance (e.g. see Mannes et al. 2014) and the exact nature of the problem (also see Hogarth 1978). The select-10 crowd allows us to investigate the sensitivity in the performance of the select crowd approach as we increase the number of included models. For select-5 crowd, the predicted choice probability for

choice problem *i* is $\hat{y}_{\text{sc5},i} = \sum_{m=1}^{M} w_m^{\text{Select-5}} \hat{y}_{m,i}$, where $w_m^{\text{Select-5}} = \begin{cases} \frac{1}{5}, \text{ if } w_m^{Akaike} \text{ among top-5;} \\ 0, \text{ otherwise.} \end{cases}$. Similarly, the prediction for select-10 crowd is $\hat{y}_{\text{sc10},i} = \sum_{m=1}^{M} w_m^{\text{Select-10}} \hat{y}_{m,i}$, where

$w_m^{\text{Select-10}} = \begin{cases} \frac{1}{10}, \text{ if } w_m^{Akaike} \text{ among top-10;} \\ 0, \text{ otherwise.} \end{cases}$.

The final model crowd was the contribution crowd. This crowd is a variant of Budescu and Chen's (2014) contribution weighted model, which leverages each individual model's unique contribution to the aggregate predictions in the training set. The contribution of model *m* is based on the comparison between the log-loss of the training set with the unweighted mean predictions of all models (denoted by $LL_{Training}$), and the log-loss with the unweighted mean predictions of all models excluding model *m* (denoted by ${}^{-m}LL_{Training}$). When a model makes a positive contribution, by decreasing the log-loss, it is given a positive weight. The magnitude of the weight is proportional to the magnitude of this unique contribution. When a model does not decrease the log-loss, it is given a zero weight (i.e. removed from the crowd). Formally, the contribution-based weight for model *m* is given as:

$$w_m^{Contribution} = \frac{\mathrm{I}\left(LL_{Training} - {}^{-m}LL_{Training} > 0\right)\exp(LL_{Training} - {}^{-m}LL_{Training})}{\sum_{k=1}^{M}\left[\mathrm{I}\left(LL_{Training} - {}^{-k}LL_{Training} > 0\right)\exp(LL_{Training} - {}^{-k}LL_{Training})\right]}, \text{ with } \sum_{i=1}^{M} w_m^{Contribution} = 1.$$

Again, in our application of this crowd, the contribution-based weights were derived solely from the training set and were applied to the test set to obtain the fully out-of-sample predicted choice probability for each choice problem *k*, written as $\hat{y}_{cc,i} = \sum_{m=1}^{M} w_m^{Contribution} \hat{y}_{m,i}$.

# Results

## Individual models

**Predictive performance.** The wide range of datasets we collected offer an ideal test-bed for evaluating the different models. As discussed above, we evaluated models' out-of-sample predictive performance with prediction error, log-loss and Brier score respectively. Figure 3 shows the models' prediction errors for gains and mixed gambles (for the results based on log-loss and Brier score, see Supplementary Figures S1-S2). For gains, the dual-systems model (Loewenstein et al. 2015) had the lowest overall prediction errors. Other close competitors included prospective reference theory (PRT, Viscussi 1989), odds-based subjective weighted utility theory (odds-based SWU, Karmarkar 1978), subjective expected utility theory (SEU, Edwards 1955) and several variants of the (cumulative) prospect theory. For mixed gambles, the leading models were two variants of cumulative prospect theory that treated gains and losses differently (Lattimore et al. 1992, Prelec 1998). This result is consistent with earlier findings that the best variant of cumulative prospect theory has a power value function and Prelec's probability weighting function (Stott 2006). Other close competitors included the transfer of attention exchange model (TAX, Birnbaum 2008), odds-based subjective weighted utility theory (Karmarkar 1978), subjective expected utility theory (Edwards 1954) and the dual-systems model (Loewenstein et al. 2015).

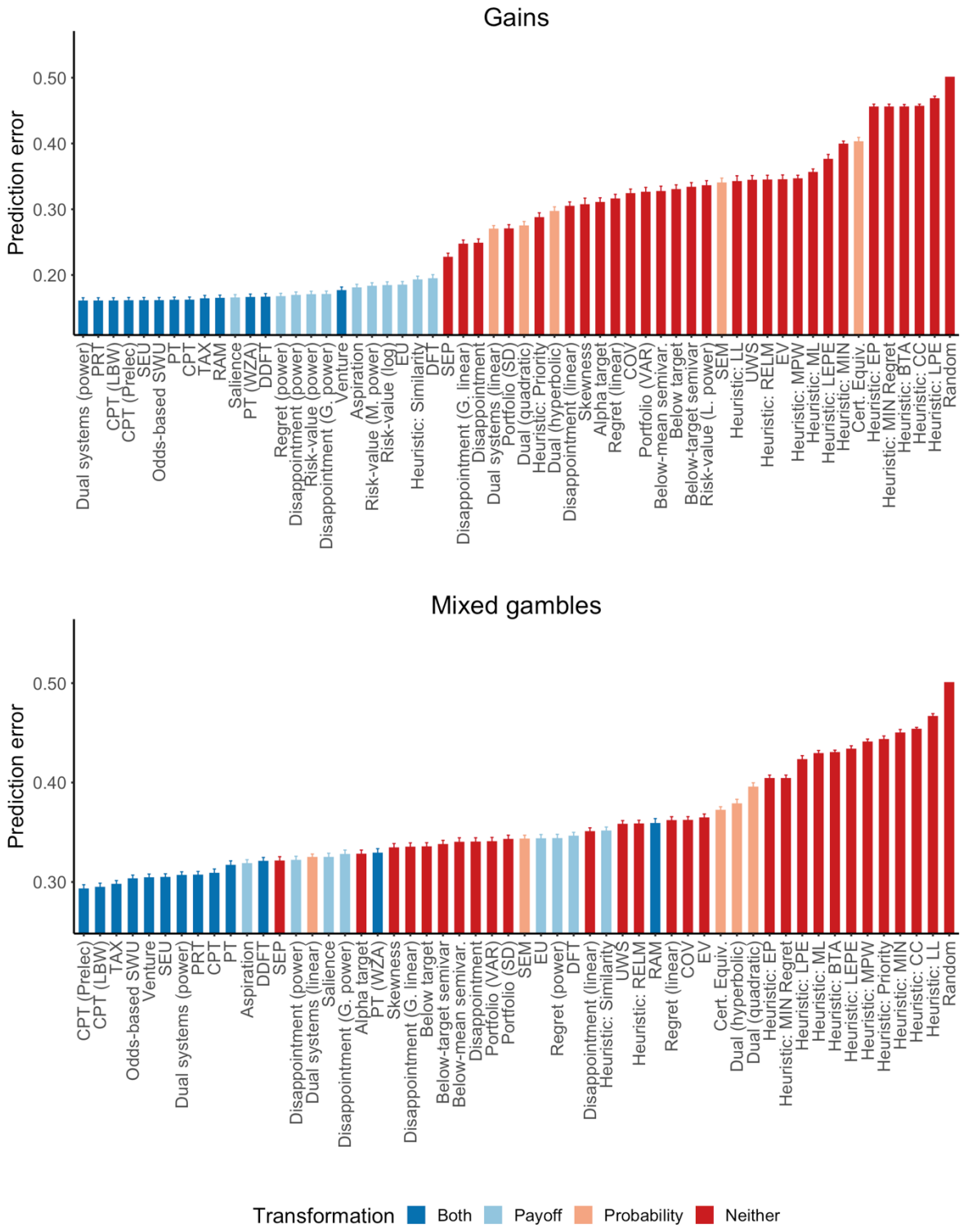


Figure 3. Individual models' mean prediction errors across participants in gains and mixed gambles. The color of the bars indicates whether or not the model transforms payoffs or probabilities. Error bars represent standard errors of predictive performance across datasets and individuals. Details and full names of the models can be found in the Appendix. (see the color figure online)

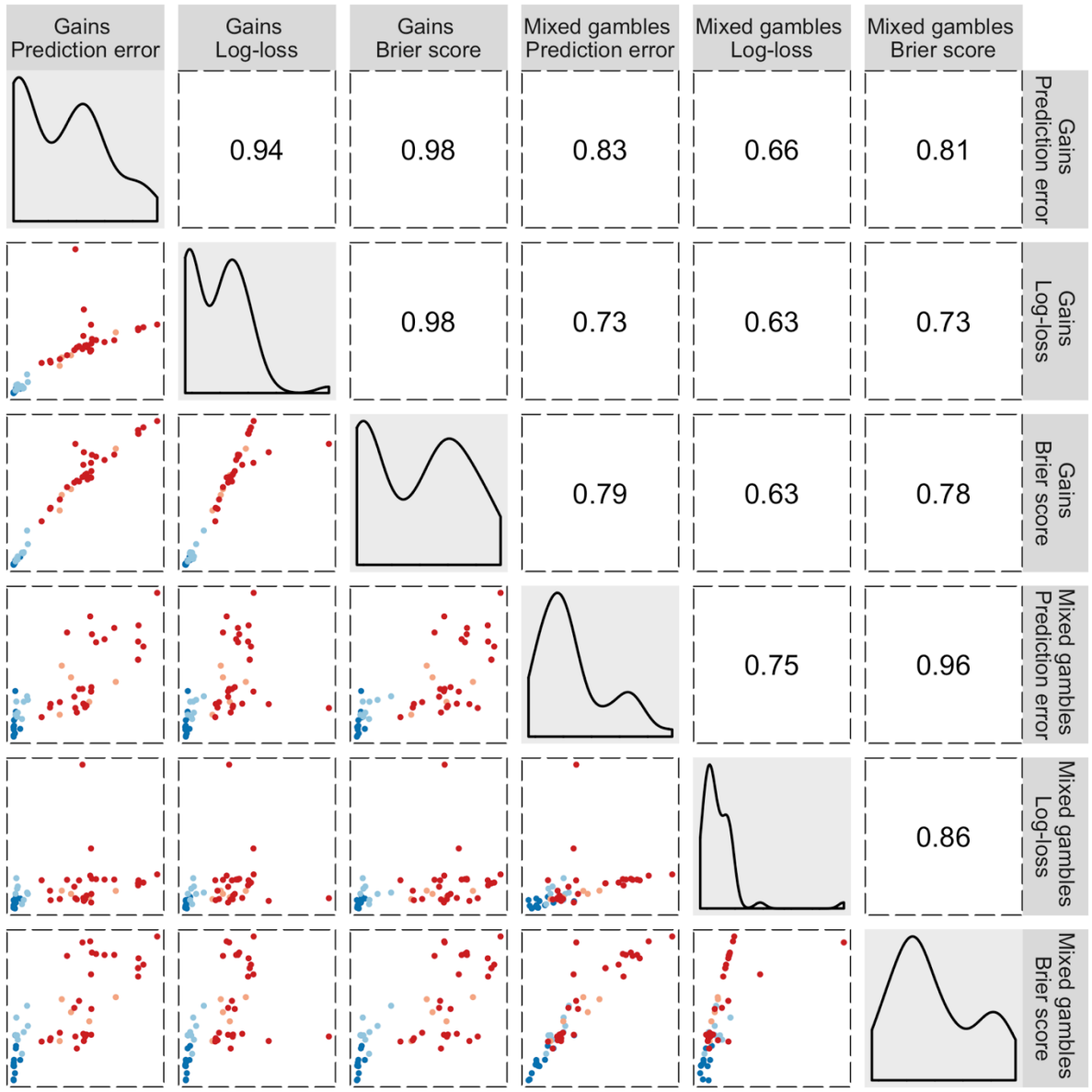


Figure 4. Summary of predictive performance in gains and mixed gambles using different loss functions. The diagonal plots the distributional densities of the predictive performance metrics. The lower cells display the scatterplots. The color of the points in the scatterplots corresponds to the color scheme in Figure 2, indicating whether the model transforms payoffs and probabilities non-linearly. The upper cells display the corresponding Spearman correlation coefficients for pairs of datasets and loss functions (all *p*s <.001). (see the color figure online)

Although the exact ranking of models in predictive performance varied with loss functions, the set of top-performing models was highly robust across loss functions. Figure 4 summarizes the similarities of model rankings across loss functions. The Spearman rank correlations between different loss functions were all above 0.94 for gains, and all above 0.75 for mixed gambles, suggesting a high consistency across loss functions. The models' relative predictive performance in the two types of datasets was also highly consistent, with high Spearman's rank correlations according to any of the three loss functions (see also Figure 4). That said, we found that the overall log-losses, Brier scores, and prediction errors were higher for mixed gambles than for gains. For example, the lowest mean prediction error achieved by a model on average across participants for gains was around 0.16, but the counterpart for mixed gambles was almost twice as much at around 0.29. Nonetheless, it is premature to conclude that the models were less accurate in predicting choices in the mixed domain than in the gain domain, as other factors in the stimulus sets, such as the EV differences, might also lead to differential predictive performance.

Going beyond the mean performance of models across participants we also examined the proportion of people for whom a certain model made best out-of-sample predictions according to the three loss functions. This analysis revealed a high degree of heterogeneity in the best-performing models at the individual level (see Supplementary Figures S3-S5 for the models' proportion of the best predictive performance across participants for the three loss functions). A large number of models accumulated no less than 2% of the best predictive performance across participants and measures for both gains and mixed gamble datasets. This indicates that although some models achieved high predictive performance on average, there was no unequivocal best-performing model on the individual level.

**Psychological mechanisms.** To assess the relative value of the nine different psychological mechanisms characterizing the considered individual models, we compared the average prediction error of all the models that make use of a mechanism with models that do not. These results are summarized in Figure 5. Our analysis suggests that payoff transformation is the most crucial psychological mechanism for improving our ability to predict risky choice---models using the payoff transformation mechanism had a prediction error of 0.17 with gains and a prediction error of 0.32 with mixed gambles, compared to prediction errors of 0.34 (gains) and 0.38 (mixed gambles) for models without payoff transformation (this corresponds to a paired-sample Cohen's *d* of 1.63 for gains and 1.50 for mixed gambles). These differences can be seen by comparing the blue and red bars and points in Figures 2 and 3. They are also clearly reflected in the bimodal distributions of predictive performance, shown on the diagonal in Figure 4: One peak of the distribution corresponds to the models with payoff transformations and the other corresponds to the models with no payoff transformations. Consistent with this result, models using the payoff transformation mechanism outperformed their counterparts that did not assume the mechanism. To name a few examples, expected utility maximization outperformed expected value maximization; the dual-systems model with a power value function outperformed the dual-systems model with a linear value function; regret theory with a power value function outperformed regret theory with a linear value function; The three relative risk-value models with payoff transformations outperformed the one relative risk-value model with no payoff transformations; The similarity model, the only heuristic model that applies payoff transformation, outperformed all other heuristics.

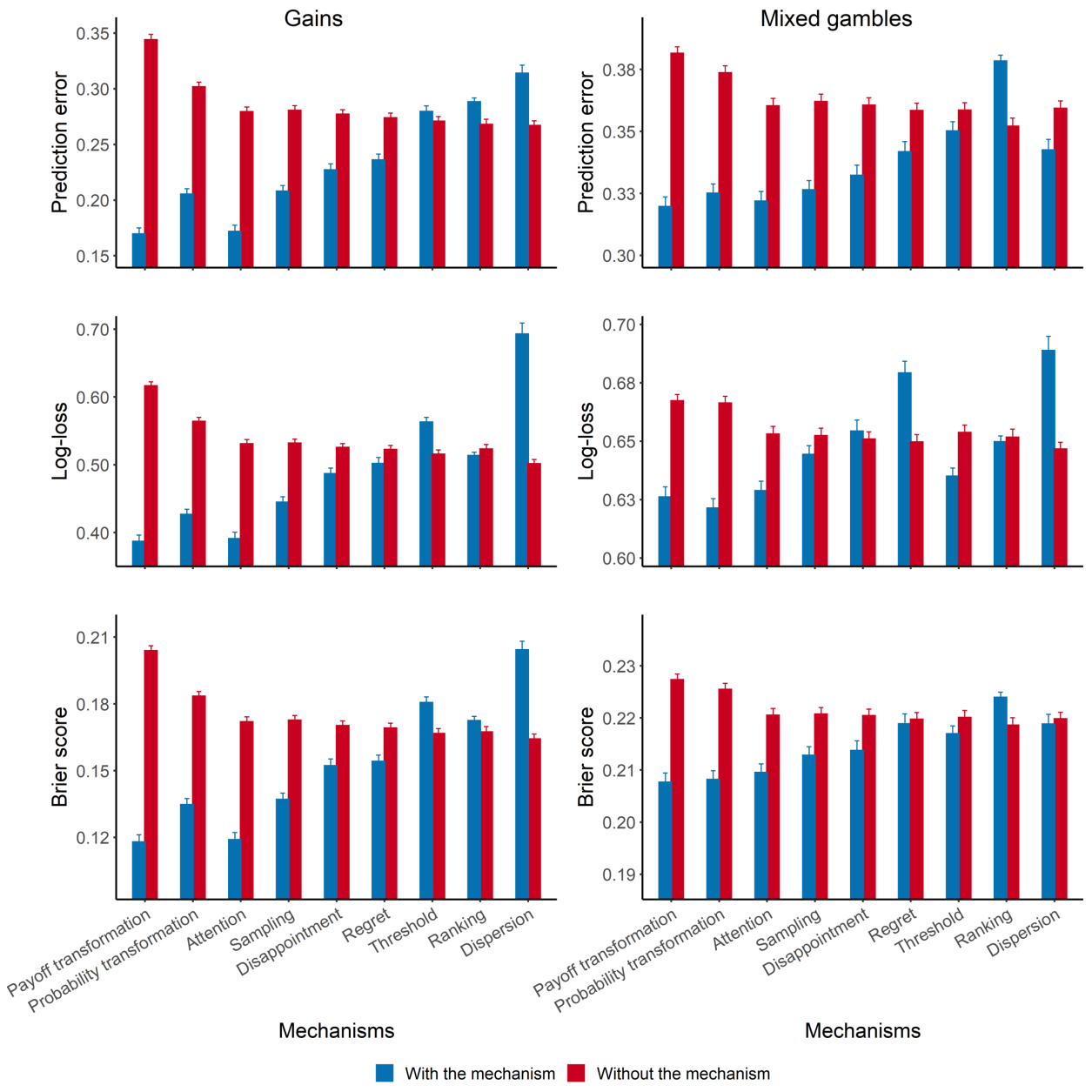


Figure 5. Mean predictive performance by models with or without certain psychological mechanism. Error bars represent standard errors of predictive performance across datasets and individuals. (see the color figure online)

Another top performing mechanism was probability transformation. Models using the probability transformation mechanism had prediction errors of 0.21 for gains and 0.33 for mixed gambles, compared to prediction errors of 0.30 (gains) and 0.37 (mixed gambles) for models without probability transformations (corresponding to a paired-sample Cohen's *d* of 2.05 for

gains and 1.27 for mixed gambles). These differences can be seen in the light red and dark blue bars and points in Figures 3 and 4. The attention, sampling, disappointment and regret mechanisms often led to favorable prediction outcomes. The attention mechanism, for example, sometimes led prediction gains comparable to those from payoff or probability transformation. The ranking, threshold and dispersion mechanisms, by contrast, did not improve prediction outcomes. Models using these mechanisms performed on average worse than models without these mechanisms. Of course, there was substantial inter-individual variability. Even the modestly or poorly performing mechanisms describe well a considerable number of individuals.

**Model crowds**

**Predictive performance**. How well did model crowds do in comparison to the individual models? To answer this question, we first compared the model crowds to the individual models that provided the best average performance across participants using the loss function scores in the test data. As shown in Figure 6, the four performance-based model crowds (i.e. the select-5, select-10, weighted and contribution crowds) outperformed all individual models and achieved the highest overall predictive performance. Here, the individual model performance metric labeled as "Aggregate best" is the best performing individual model in aggregate as in Figure 3 and Supplementary Figures S1-S2 depending on the loss function implemented (later on we examine individual model performance with an alternate metric, labeled "Training-contingent".) This pattern was true for both gains and mixed gambles. Although the naïve crowd did not outperform all the individual models, it still surpassed a large majority of them, lagging behind only a few individual models (implying that it still demonstrated the wisdom of the crowds, see e.g. Davis-Stober et al. 2014). Overall, the model crowd approach can improve overall predictive performance in an out-of-sample manner. This was especially true when the aggregation

strategies assigned larger weights to models that perform better at the training stage (as in the select and weighted crowds) or when aggregation strategies leverage each individual model's unique strength in predicting choice behavior (as in the contribution crowd). Notably, the advantage of model crowds over individual models was robust across all loss functions and was even more pronounced with the probabilistic loss functions (i.e. log-loss and Brier score), that were inherently more sensitive to continuous model predictions.

Not only did model crowds have better average performance across participants, they also made better predictions for a majority of participants when compared to the best performing individual models. The rows labeled "Aggregate best" in Table 3 show the proportion of participants for whom a model crowd made better predictions than the best individual models using the various loss functions on the test data. As can be seen here, with log-loss as the loss function, the contribution crowd made better predictions than the best individual model (which was CPT with Prelec's probability weighting function) for 72% of the participants in the *gains* datasets. In the mixed gambles datasets, the number went up to 88%. These patterns were robust to different loss functions, as well as different model crowds. The only exception, however, was the naïve crowd for the *gains* datasets. The naïve crowd did not make better predictions for a majority of participants when compared with the best individual models in gains. Its average predictive performance was also inferior to the best-performing individual models (see Figure 5), suggesting that in risky choice considering different models as equally valid in the model crowd may not be the best way to leverage the collective wisdom of individual models.

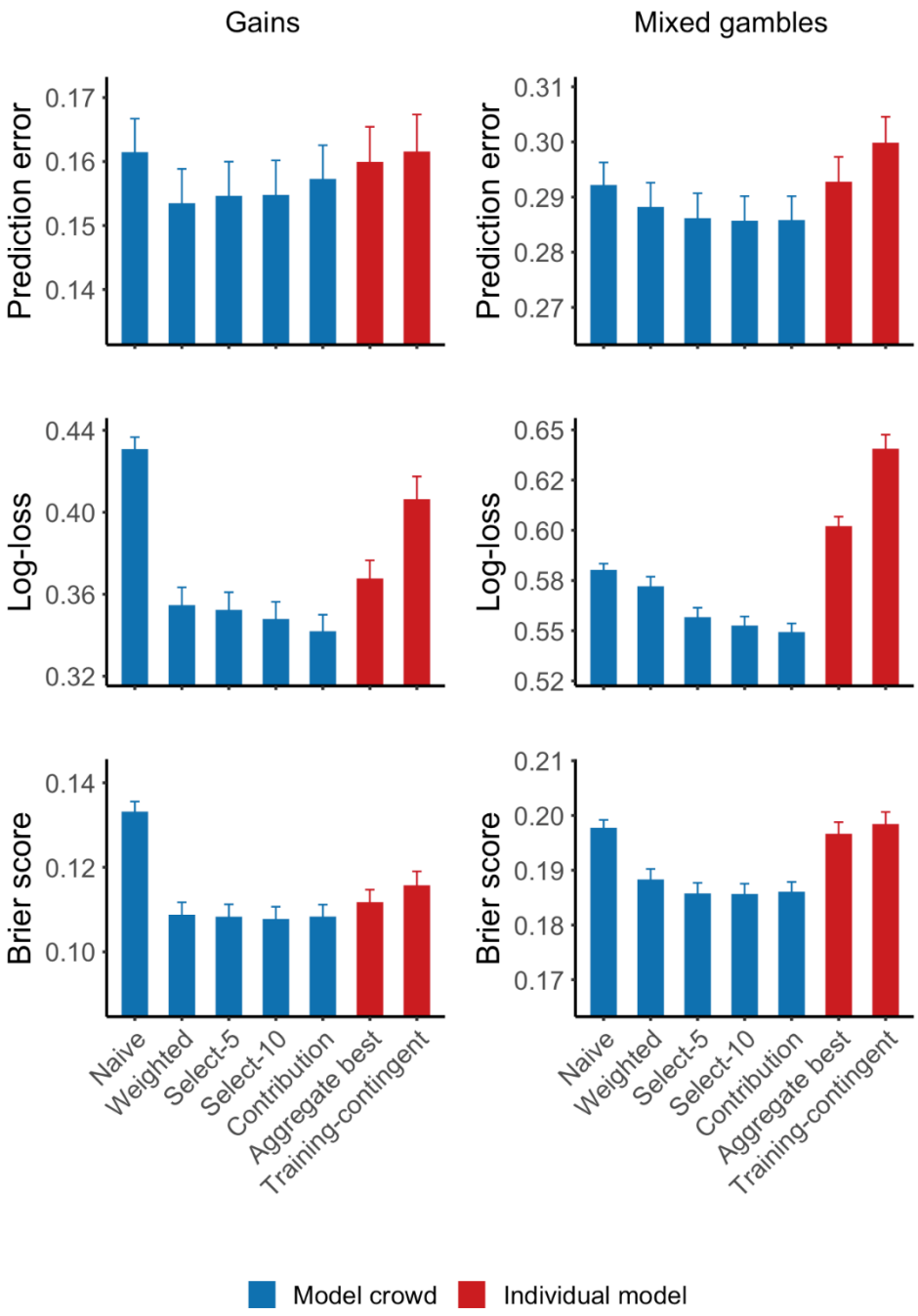


Figure 6. Predictive performance of model crowds and best individual models. Error bars represent standard errors of predictive performance across datasets and individuals. (see the color figure online)

Table 3. Proportion of participants for whom the model crowds provide better predictive performance than the best individual models.

| Gains | | | | | | |
|---|---|---|---|---|---|---|
| Individual model | Loss function | Model crowds | | | | |
| | | Naïve | Weighted | Select-5 | Select-10 | Contribution |
| Aggregate best | Prediction error | 0.44 | 0.51 | 0.52 | 0.54 | 0.49 |
| Aggregate best | Log-loss | 0.22 | 0.51 | 0.56 | 0.65 | 0.72 |
| Aggregate best | Brier score | 0.16 | 0.55 | 0.6 | 0.68 | 0.64 |
| Training-contingent | Prediction error | 0.50 | 0.64 | 0.61 | 0.61 | 0.57 |
| Training-contingent | Log-loss | 0.34 | 0.77 | 0.78 | 0.76 | 0.76 |
| Training-contingent | Brier score | 0.25 | 0.71 | 0.72 | 0.71 | 0.66 |
| Mixed gambles | | | | | | |
| Individual model | Loss function | Model crowds | | | | |
| | | Naïve | Weighted | Select-5 | Select-10 | Contribution |
| Aggregate best | Prediction error | 0.53 | 0.54 | 0.56 | 0.58 | 0.58 |
| Aggregate best | Log-loss | 0.65 | 0.59 | 0.71 | 0.80 | 0.88 |
| Aggregate best | Brier score | 0.51 | 0.6 | 0.68 | 0.74 | 0.73 |
| Training-contingent | Prediction error | 0.57 | 0.63 | 0.61 | 0.62 | 0.60 |
| Training-contingent | Log-loss | 0.66 | 0.79 | 0.80 | 0.77 | 0.77 |
| Training-contingent | Brier score | 0.53 | 0.70 | 0.72 | 0.69 | 0.67 |

*Note*. When the two models provide the best out-of-sample predictive performance (especially with prediction error as the loss function), they split the best-performance count.

Note that the model crowds weigh the different models based on their performance in the training data. This allows them to flexibly identify best performing models (in the training set) for each individual and use these models to make predictions. Thus, the specific weighting scheme used by a particular model crowd varies across individuals. It could be this flexibility, rather than crowd wisdom, that results in the better performance of model crowds in Figure 5 and Table 3. To ensure that this was not the case, we contrasted our model crowd predictions with those of individual models with the same type of flexibility. This was done with an approach that

flexibly paired each individual participant, at every split in the 10-fold cross-validation, with the best-performing model in the individual's training set, evaluated using the Akaike information criterion. This *training-contingent* algorithm can be seen as a select crowd with only the most promising model included (corresponding to a select-1 crowd). We then used this training-contingent model, to make predictions in the test sets, for each participant for every split, and evaluated its out-of-sample predictive performance with the above loss functions. The results of this analysis are shown in the "Training-contingent" bars in Figure 6 and rows labeled "Training-contingent" in Table 3. Here we can see that the training-contingent approach actually reduces predictive performance relative to the fixed individual models that provide the best predictions across participants (likely due to the high variance of this algorithm). Thus, the advantage of model crowds over individual models cannot be attributed to their flexibility in identifying the best individual model in a particular split of the training data. Rather, it is likely due to crowd wisdom, which exploits the complementarities of different models that make up the crowd.

Finally, the four performance-based model crowds had an obvious advantage over the naïve crowd that treated all individual models as equally valid---performance-based crowds can leverage the differential predictive power of individual models. The four performance-based model crowds achieved roughly the same predictive ability with prediction error and Brier score as loss functions. However, with log-loss as the loss function, the contribution crowd tended to provide the best overall predictive performance for both gains and mixed gambles (see Figure 6). Our analysis of model weights in the next section will unpack potential causes of the contribution crowd' superior predictive performance. The historical analysis that follows also illustrates an additional strength of the contribution crowd: it can successfully aggregate model predictions regardless of the number and performance variance of the models present in the

model pool. This will become apparent when looking at historical time windows where low-performing models are overrepresented (see Figure 9 for more details).

**Weights in model crowds.** The distribution of model weights differed across model crowds. To examine this, we calculated for each participant a measure of weight dispersion in each model crowd using the Gini coefficient, a canonical measure of dispersion and inequality in distributions. If all weights concentrate on one single model, the Gini coefficient will be 1, meaning that there is a minimal amount of dispersion. By contrast, in the naïve crowd where each model receives an equal weight, we have a Gini coefficient of 0, meaning a maximal amount of dispersion. The dispersion of model weights in the four performance-based crowds lies in between. As in Figure 7, select and weighted crowds mostly concentrate on a few models with the mean Gini coefficients between 0.8 and 0.9 while the contribution crowd strikes for a more balanced dispersion of model weights, with Gini coefficients around 0.5.

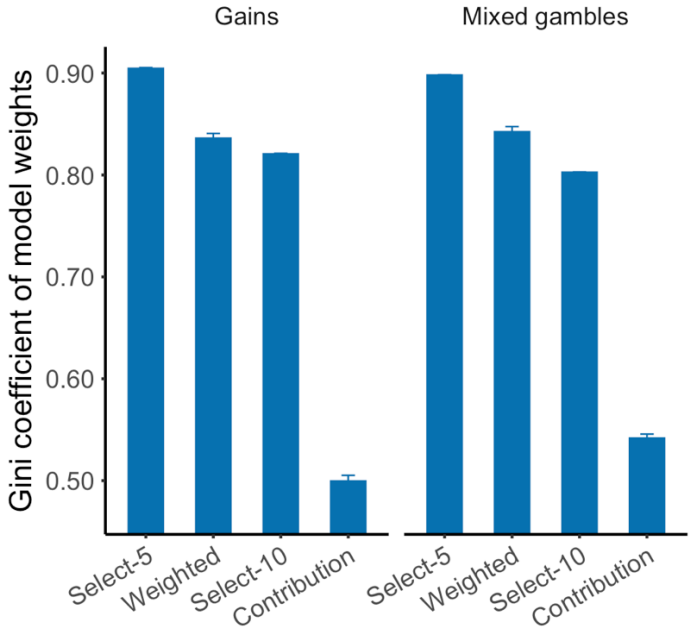


Figure 7. Dispersion of model weights in model crowds (measured with Gini coefficient). Error bars represent standard errors of Gini coefficients across datasets and individuals.

The success of model crowds can be also understood in terms of the distribution of model weights. Model crowds trade-off between assigning larger weights on the best performing individual models, and hedging their bets by dispersing the weights more across different models (Davis-Stober et al. 2014, Müller-Trede et al. 2017). The training-contingent model (which would correspond to the select-1 crowd) and the naïve crowd represent two boundary solutions to this trade-off. The former goes all-in and adopts the prediction of the best performing model in the training set (leading to a Gini coefficient of 1), while the latter is maximally diverse and unconditionally averages the predictions of all the models, regardless of model performance in the training set (leading to a Gini coefficient of 0). Yet, as shown in Figure 6, neither of the two boundary solutions performed as well as the four performance-based model crowds, the latter of which struck for a more balanced distribution of model weights. The contribution crowd, in particular, has been the best performing model in many of our tests using probabilistic loss functions in both gains and mixed gambles. This result indicates that good crowd solutions may, in fact, leverage a quite diverse crowd of models (i.e. Gini coefficient 0.5 for the contribution crowd, also see Hong and Page 2004, Lamberson and Page 2011).

The weights assigned to individual models in the model crowds also allowed us to measure the degree to which different models contributed to the crowd predictions. Figure 8 displays the weight each individual model received in the contribution crowd, the performance-based crowd that makes the best use of model diversity (see Supplementary Figures S6-S8 for selected and weighted crowds). As expected, the top contributors were often the models that did very well in the individual model comparison. Moreover, the fact that all models (except the random model) made non-zero contributions in both the gains and the mixed datasets indicates that each existing model captures some unique features of choice behavior.

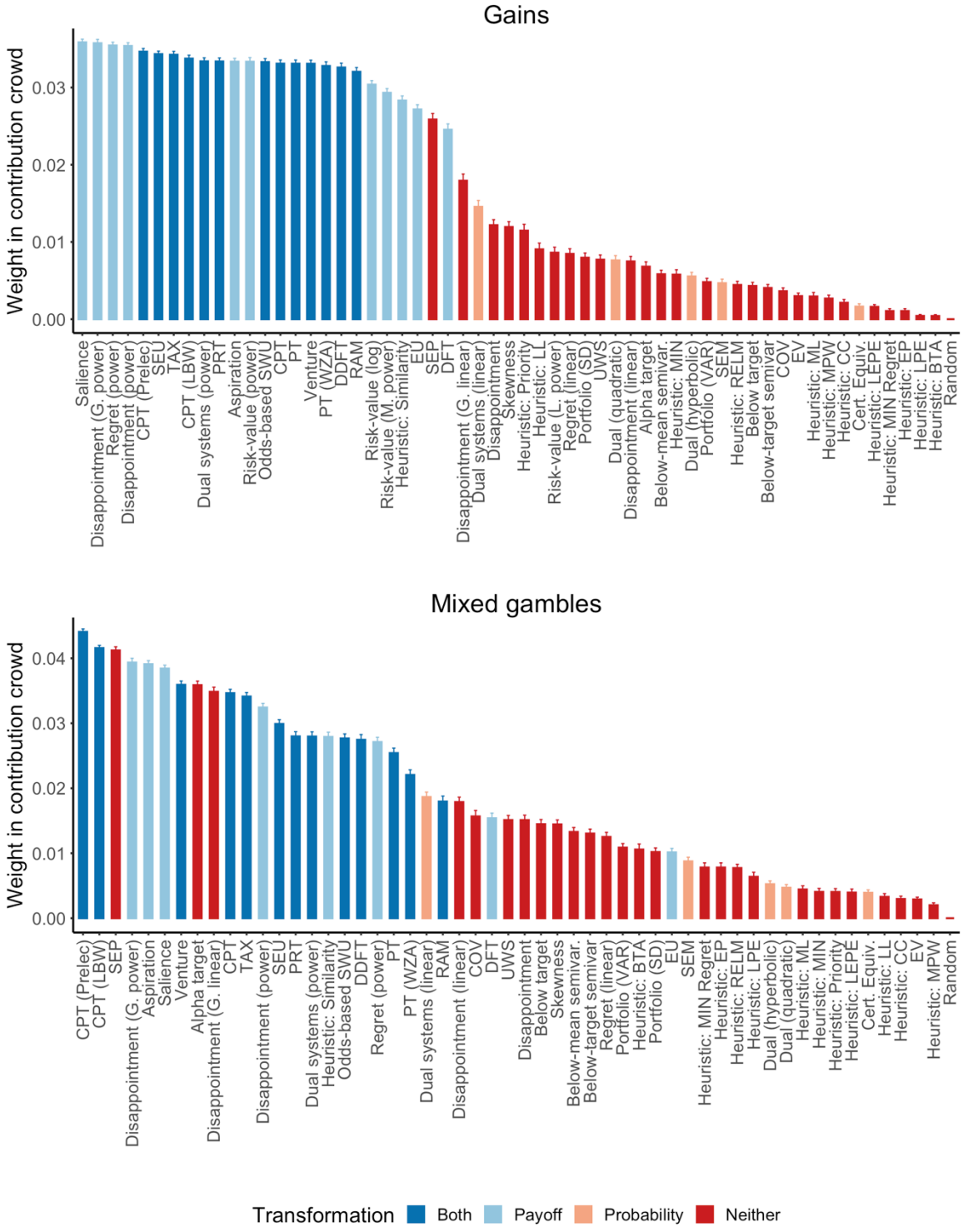


Figure 8. Model weights in the contribution crowd. The color of the bars indicates whether or not the model transforms payoffs and probabilities non-linearly. Error bars represent standard errors of model weights across datasets and individuals. (see the color figure online)

**Historical trends**

We also evaluated how predictive accuracy of risky decision models evolved historically. Our historical analysis started from the year 1950, at which point there were only three models (the baseline random model, expected value theory and expected utility theory), and extended until the year of 2018, at which point, all the models involved in the current analysis had been published. We first evaluated performance of the best-performing individual model at each point in time. This has been relatively stable over most of the historical timeline. For gains, expected utility theory was the best performing model before 1950, until the advent of subjective expected utility (SEU, Edwards 1954). Afterwards, there were minor improvements in predictive accuracy with odd-based subjective weighted utility theory formulated by Karmarkar (1978), prospective reference theory (Viscussi 1989), two variants of cumulative prospect theory (Lattimore et al. 1992, Prelec 1998), and the dual-systems model (Loewenstein et al. 2015), depending on the loss function implemented (see Figure 9).

For mixed gambles, expected utility theory was also the best model at the beginning. However soon it was supplanted by portfolio theory (Markowitz 1952), and then again by subjective expected utility theory, which led to a big leap in predictive performance. This model remained the best-performing model for more than two decades until odds-based subjective weighted utility theory was introduced. A significant historical leap came with the introduction of models that treated gains and losses differently, such as cumulative prospect theory and the transfer of attention exchange model (TAX, Birnbaum 2008). Again, there are minor differences based on the specific loss function used.

We also evaluated the performance of our model crowds at each historical time point. As can be seen in Figure 8, the contribution crowd outperformed the best individual model available for nearly all time points (regardless of the number and the composition of models involved in the crowd). This was true for both gains and mixed gambles and the advantage of the contribution crowd was even more pronounced for mixed gambles. These results again show that aggregation algorithms that successfully exploit each model's strength in predicting idiosyncratic individual-level data, while hedging their bets across different models, can reliably predict risky choice behavior better than individual models.

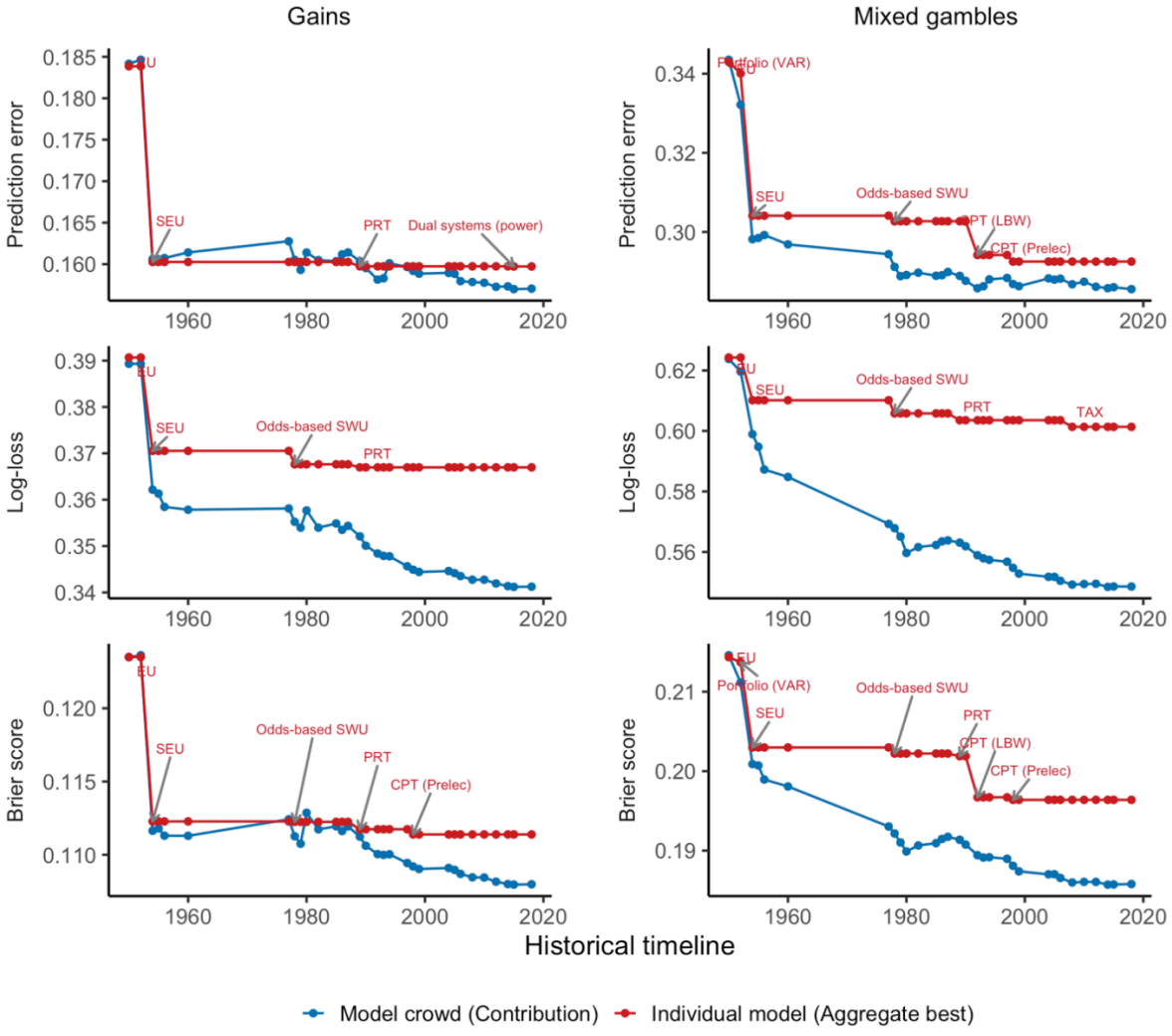

Figure 9. The historical evolution of predictive performance. For the model crowds, all models formulated up to the year reported on the *x*-axis were used to calculate the crowd predictions for that time point. CPT: cumulative prospect theory; EU: expected utility; PRT: prospective reference theory; SEU: subjective expected utility; SWU: subjective weighted utility; TAX: transfer of attention exchange. (see the color figure online)

Model crowds other than the contribution crowd showed slightly different patterns (the historical trends of all model crowds can be found in Supplementary Figure S9). These crowds underperformed in early time periods, when only a limited number of models were available. This was especially the case for the naïve and select crowds, which assigned equal weights to all models or to subsets of models used in the crowd. The weighted crowd was more robust to the effects of small crowds. However, with the introduction of newer behavioral models, the performance of model crowds greatly improved, and all model crowds outperform the best individual model from the 1970s onwards. Overall, while being able to leverage crowd wisdom, the select and weighted crowds appeared to be more sensitive to the composition of the model pool than the contribution crowd.

**Psychological mechanisms in model crowds**

Not only can model crowds improve performance, but can be also used to assess the relative importance of different psychological mechanisms for the study of risky choice. The first way to achieve this is to evaluate the average weights of models that have a specific mechanism and compare them to the average weights of models that do not have this mechanism. This analysis reveals that models with payoff transformation have much larger weight in the contribution crowd than models without it. The difference in average weights was pronounced for models with the attention, probability transformation and disappointment mechanisms, and moderate for the sampling and regret mechanisms (Figure 10). By contrast, the average weights

of models with the threshold, ranking and dispersion mechanisms are lower than those of models that do not have these mechanisms.

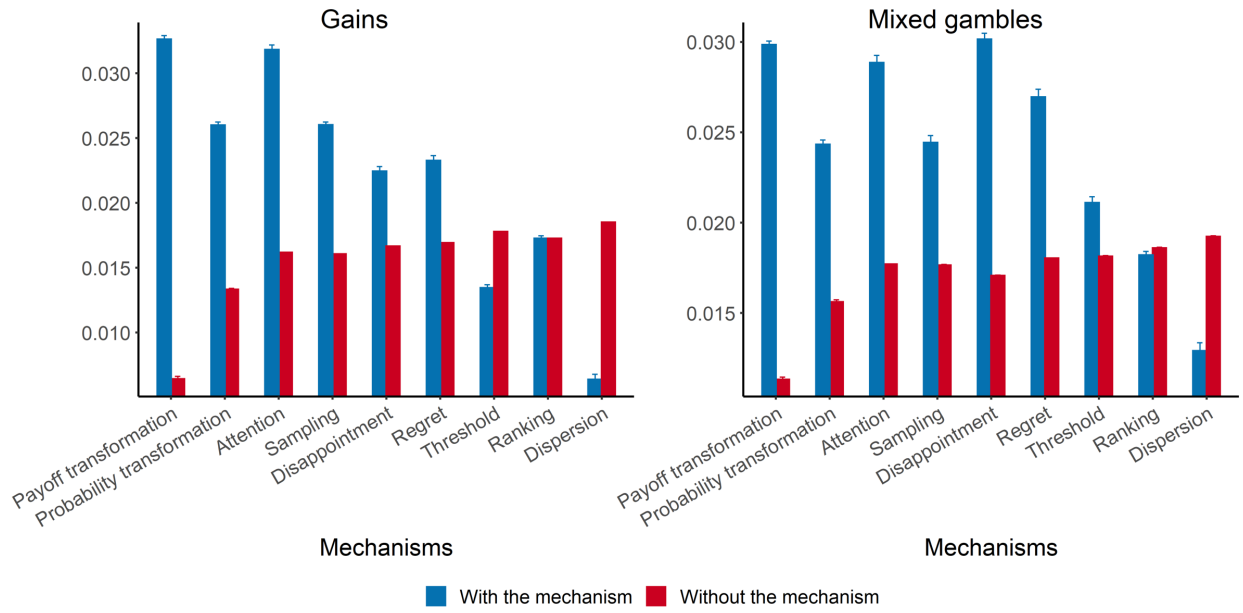


Figure 10. Mean weights of the models that use a mechanism as compared to the models that do not use it in the contribution crowd. Error bars represent standard errors of model weights across datasets and individuals. (see the color figure online)

A second approach to assess the relative impact of each of the nine psychological mechanisms on prediction is to remove all the models that involve the psychological mechanism from the contribution crowd (removing mechanisms one at a time). This is a process like the historical analysis of the contribution crowd (see the dotted blue line in Figure 9), but this time models are filtered at the mechanism level. As in Figure 11, removing models with payoff transformation substantially increased the prediction error in the contribution crowd, compared with the model crowds using all the models as in Figure 5. The same happened, but to a lesser extent, when models belonging to the probability transformation mechanism were removed. By contrast, removing any other psychological mechanism did not appear to significantly influence the crowd's predictive performance. The results from both these analyses are largely consistent

with the individual level analysis of the different mechanisms, which identified payoff and probability transformations as the two most important mechanisms for improving predictive performance, followed by attention, sampling, disappointment and regret.

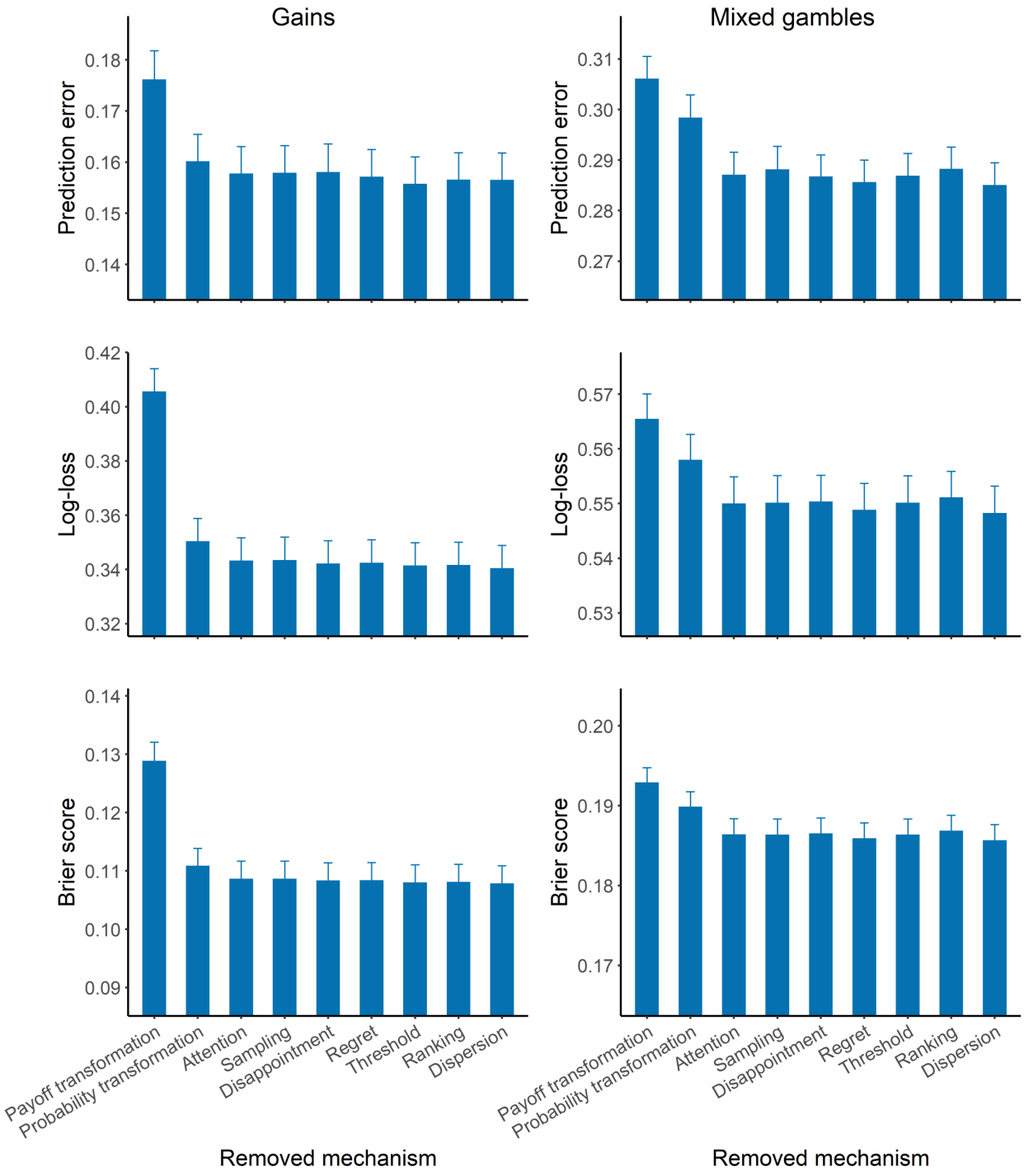

Figure 11. Predictive performance of the contribution crowd using the subset of individual models that do not involve the mechanism on the *x*-axis. Error bars represent standard errors of predictive performance across datasets and individuals.

**Impact of experimental designs**

Finally, we ran a sensitivity analysis by examining the extent to which the results varied across datasets. The relative rank of models' prediction errors was highly consistent across datasets, with high Spearman rank correlation $\rho$ (median = 0.88, mean = 0.90). This suggests that there is converging evidence across datasets with regards to the models' predictive performance. The same holds true when considering the contribution of different mechanisms to predictive performance. To further bolster this point, we calculated a paired-sample Cohen's *d* for each mechanism, by comparing the average predictive performance of models that used the mechanism to that of models that did not use the mechanism, for each dataset. This is shown in Figure 12, which displays the distribution of Cohen's *d* across datasets using different loss functions. Mechanisms such as payoff transformation, probability transformation, attention, sampling, disappointment and regret reliably boost predictive performance across datasets for gains while threshold, ranking and dispersion show ambiguous patterns. The patterns for mixed gambles were similar, except that the disappointment and regret mechanisms became less productive for mixed gambles than for gains. The predictive advantage of model crowds over individual models also persisted in 13 out of the 19 datasets (68.4%), even if we allowed each dataset to be paired with its own best performing individual model. Overall, the results discussed in the above sections were corroborated across most datasets.

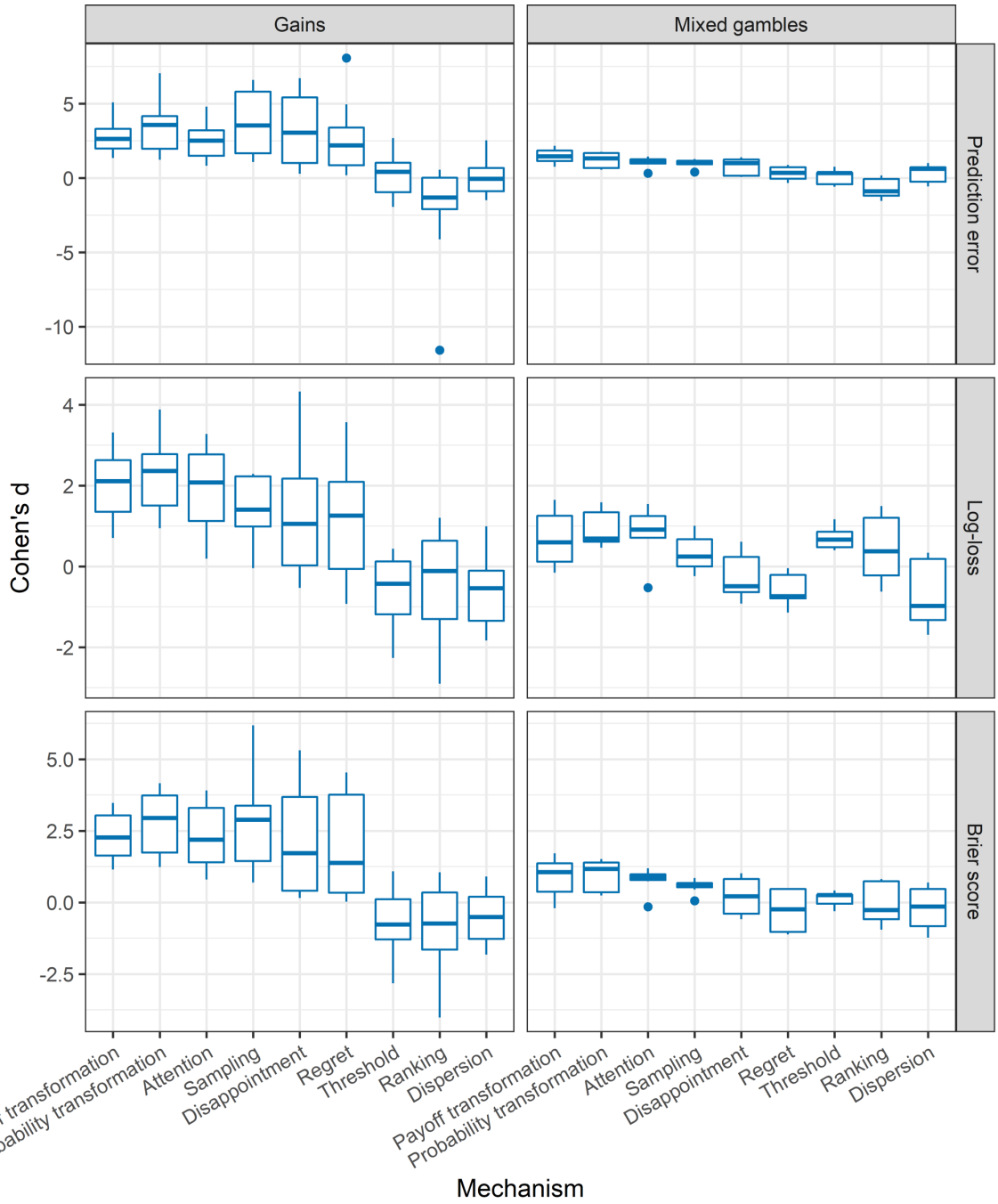


Figure 12. Boxplot distribution of Cohen's *d* of each psychological mechanism across datasets. For each dataset, Cohen's *d* was calculated by comparing the mean predictive performance of the models that involve a mechanism with that of the models that does not involve the mechanism. Each boxplot is composed of 12 Cohen's *d* values in gains (corresponding to the 12 *gains* datasets) and 7 Cohen's *d* values in mixed gambles (corresponding to the 7 *mixed* datasets).

Yet, there were still small differences that can be attributed to specific experimental designs. This can be in part seen in the distribution of Cohen's *d* in Figure 12. To further illustrate this, we map out the datasets on a 2-dimensional plane using multidimensional scaling (MDS). Specifically, we calculated a Spearman rank correlation $\rho$ between each pair of datasets in terms of the models' predictive performance and then used 1-$\rho$ as the distance measure for MDS. As shown in Figure 13, there appears to be a gap between gains and mixed-gamble datasets. This gap is largely driven by manually created datasets with one nonzero branch in the choice (i.e. Stewart et al. 2015 [SRH15], Stewart et al. 2016 [SHM16]), and does not appear with datasets that involved randomly generated gambles involving two nonzero branches. The gain datasets that have two nonzero branches (i.e. Fiedler and Glöckner 2012 [FG12], Rieskamp 2008 [Rieskamp08]) are closer to the mixed-gamble datasets (which involve randomly generated gambles and have two nonzero branches in the gamble) than to the gain datasets with manually created one-branch gambles.

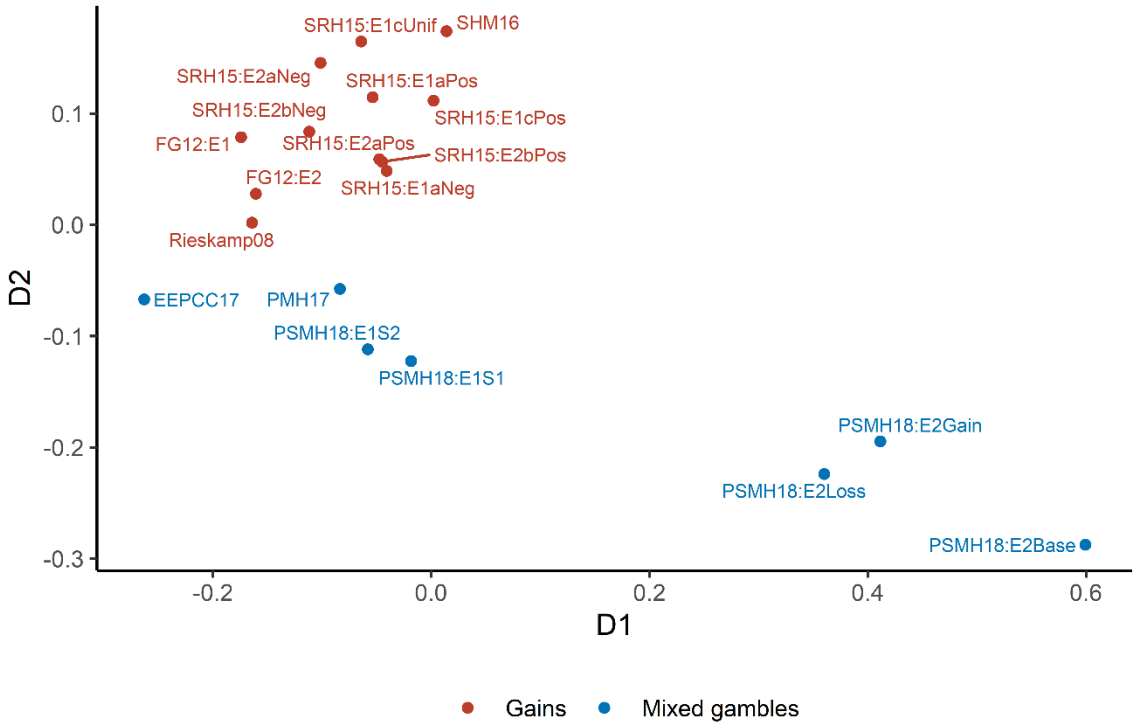


Figure 13. Two-dimensional solution when applying multidimensional scaling on the datasets using 1-ρ (where ρ corresponds to Spearman rank correlation) as the distance. (see the color figure online)

Among the mixed-gamble datasets, there was a notable difference between datasets from Experiment 2 of Pachur et al. (2018) and others. Specifically, in Pachur et al.'s (2018, Experiment 2) datasets, heuristic models such as the better-than-average heuristic, the minimax regret heuristic and the equiprobable heuristic (Thorgate 1980) performed reasonably well, while in other datasets the same heuristic models predicted poorly. This is likely because the design of the stimuli was favorable to these heuristic models. For example, some choice items in this experiment were simply rejecting/accepting a gamble with equal odds of winning and losing. For such items, heuristic models such as the equiprobable heuristic can mimic many utility-maximizing models while being more parsimonious. This analysis reveals that although the results are remarkably stable across datasets, specific design choices may still have an impact on the relative performance of different decision models.

## Discussion

For several decades, researchers have been searching for a model to describe and explain risky choice. This effort has resulted in dozens of mathematically distinct models that have their origins in several scientific disciplines. Yet, there has been little consensus with regards to the state-of-the-art in terms of predictive or descriptive performance -- different papers often compare model performance on different datasets and assess the performance of only small subsets of "rival" models. Different models are commonly seen as competitors, with the success of one model undermining other theoretical accounts. As things stand, it is hard to assess the accumulated wisdom on risky choice from a decades-long multidisciplinary research endeavor.

Our paper hopes to address some of these issues using a very large-scale model comparison. For this comparison we complied a panel of 58 existing risky choice models and compared their performance using a comprehensive test-bed of 19 existing risky choice datasets that involved over 800 participants. Further, drawing on insights from the wisdom-of-crowds literature, we tested the predictions of model crowds that aggregate the predictions of individual models.

This analysis uncovered a number of novel results regarding the predictive potential of risky choice models and model crowds. First, the best performing models fell into the category of non-expected utility theories and were often variants of prospect theory with both non-linear transformations of payoffs and probabilities (Edwards 1955, Karmarkar 1978, Lattimore et al. 1992, Prelec 1998). Other models such as the dual-systems model (Loewenstein et al. 2015) and transfer of attention exchange model (TAX, Birnbaum 2008) also made good predictions. Importantly, there was substantially individual-level variability, and most models did well for at least a few participants. Second, model crowds, and especially crowds that wisely leveraged the

diversity of the model pool, substantially improved predictions across different measures and in most datasets. Model crowds also provided a novel quantitative methodology for tracking historical accumulation of knowledge and for identifying key psychological mechanisms in risky choice modeling. Finally, the vast number of datasets allowed us to examine the important yet elusive impact of experimental designs on model selection. Although model performance strongly correlated across different datasets, our exploratory analysis revealed that design choices such as gains vs. mixed gambles, randomly generated vs. manually curated items and one-branch vs. two-branch gambles, moderated to some degree the relative performance of the competing models.

**The predictive power of model crowds**

Human behavior is highly idiosyncratic such that the model that works well for one individual may do poorly for another. Furthermore, the decision rules that guide choice are inherently noisy, reflecting fluctuations in various cognitive, affective and contextual variables (Bhatia and Loomes 2017, Busemeyer and Townsend 1993). People may also switch between decision rules depending on the nature of decision making problem at hand, a behavior commonly referred to as strategy selection (Payne et al. 1988, Lieder and Griffiths 2017). For example, they may rely on utility maximization in some problems, but switch to heuristics in others. Alternatively, different strategies may even interact in a single choice problem, a phenomenon commonly referred as strategy blending (see Erickson and Kruschke 1998, Plonksy et al. 2017, Herzog and von Helversen 2018). Thus, trying to identify the one individual model that people use, might not be the most productive approach when we want to predict people's behavior.

The model crowd approach outlined in this paper seeks to accommodate a multitude of models that take diverse theoretical perspectives (for a similar take on the social sciences, see Smaldino 2017). Of course, we do not assume that decision makers deliberate exactly like our model crowds. Rather, model crowds allow researchers to approximate the diversity of human mental processes, which results in improved performance. Indeed, the best performing model in much of our analysis was the contribution crowd, which is a crowd model that relies on considerable model diversity, as assessed by the dispersion of the model weight vector.

The model crowds' superior predictive performance can also be understood in terms of the bias-variance trade-off (see Geman et al. 1992, Gigerenzer and Brighton 2009). The total error of predictive models in machine learning, statistics and cognitive science, can be decomposed into three error components: bias, variance, and irreducible noise. Model crowds (or ensembles in machine learning) drastically reduce the variance component of prediction error (Breiman 1996), thereby improving overall prediction. This is especially the case in the presence of inter-individual variability, as in our datasets. Going with the single best-performing model would lead to good performance if we were able to identify the right model for each individual ahead of time. However, matching an individual to the best performing model is a challenging problem (Davis-Stober et al. 2014). This is clearly illustrated in the case of the training-contingent model: simply selecting the best performing model in the training-set often does not lead to the best predictions in the test set. Although this approach scores low on bias, it suffers from high variance, as it is sensitive to the specific sample that was used to find the best-performing model. The model crowds strike a much better balance on the bias-variance trade-off by substantially reducing the risk of going all in on a single model without putting much (or any) weight on others (see the supplement of Analytis et al. 2018 for further discussion). In sum,

model crowds provide a statistically reliable approach that allows to capture inter-individual variability in risky choice.

There are similar successful applications that harness collective model wisdom in decision analysis. Scheibehenne et al. (2013), for example, formulate the metaphor of the heuristic toolbox in a hierarchical Bayesian framework, and show that by incorporating multiple heuristics, the toolbox explains behavioral data better than a single heuristic. Another example is recent risky choice prediction competitions, in which researchers were provided with ample training data and were challenged to develop new modeling approaches or existing behavioral or machine learning models that could predict the proportion of people making a risky choice in a held-out test set. The most successful models in these competitions were ensembles or hybrid models encompassing insights from several decision strategies (Erev et al. 2010, 2017, Plonsky et al. 2017). Our paper extends this line of analysis by including all previously proposed risky choice models that can be translated to computer code, generating predictions at the individual level and using them as elements to construct model crowds.

**A historical perspective**

Our framework also provides a historical window onto the evolution of the field of risky choice modeling, from the axiomatization of expected utility theory by von Neumann and Morgenstern to the sophisticated behavioral models of the present day. We find an increase in the rate at which new models have been introduced in the pool of available models, but diminishing returns in overall predictive accuracy. In fact, for some measures and datasets, it has been more than a decade since a new model has outperformed the best performing model up to that point. A different picture emerges when we look at crowd models: Instead of a stagnating field we see a field with rapid improvement and continual progress. The introduction of new

models adds to our collective ability at predicting risky choices across measures for both gains and mixed gambles.

We can also use our framework to look at the importance of specific models over time. Prospect theory, arguably the most prominent behavioral decision model, performed only modestly for both gains and mixed gambles in our model comparison. In fact, subjective expected utility, formulated by Edwards in 1954, and odds-based subjective weighted utility published by Karmarkar in 1978, almost contemporaneously with prospect theory, outperformed prospect theory in terms of predictive performance both for gains and mixed gambles. These models share their core assumptions with prospect theory (such as non-linear transformations of both payoffs and probabilities). Nonetheless, models that were later derived from prospect theory outperformed these earlier models and were often among the top contestants. This is especially true for mixed gambles, in which models derived from prospect theory excel due to the assumptions of loss aversion and differential probability weights for gains and losses. Thus, although the original prospect theory was never historically the best performing model, the new concepts that were introduced in the field with prospect theory had a long-lasting impact and eventually led to improvements in our ability to predict risky choices. This is a common motif in science, ideas need to be further refined and elaborated on to reach their full potential.

**Promising psychological mechanisms**

Our large-scale model comparisons identified the psychological mechanisms that yield good predictive performance in risky choice. Payoff transformation was by far the most important mechanism, followed by probability transformation. Thus, it comes as no surprise that best performing models, including the variants of prospect theory mentioned above, fall into the category of non-expected utility theories, which use some non-linear function to transform crude

payoffs to subjective values, and in addition transform objective probabilities to subjective probabilities. This pattern emerges in both gains and mixed gambles. Payoff transformation, in particular, always improved model performance, regardless of other model characteristics involved. The payoff and probability transformation mechanisms were followed by the attention, sampling, disappointment and regret mechanisms. These three mechanisms have been used often in recent years and show some promise in their potential to improve our ability to predict risky choice, especially when combined with payoff and probability transforms.

Results on the relative importance of different psychological mechanisms for prediction were replicated in model crowds. Specifically, we tested the predictive value of different psychological mechanisms (i) using the weights in model crowds and (ii) by removing all the models using a mechanism from the contribution crowd. Once again, subjective payoff and subjective probability transformation mechanisms stood out as key mechanisms for improving predictive performance in risky choice. Predictive performance dropped substantially when models using these mechanisms were removed from the contribution crowd. The value of other mechanisms is more modest, but our analysis of average weights in model crowds suggests that it is always non-negligible, especially for models using the attention, sampling and disappointment mechanisms. That said, it is important to note that the exclusion of each of these individual mechanisms from the crowd did not substantially impact performance. Unlike payoff and probability transformation, the predictions of the individual attention, sampling and disappointment mechanisms can be mimicked by a combination of other psychological mechanisms.

Using the models in their original forms, our psychological mechanism analysis reflects each mechanism’s contribution in the research enterprise of risky decision making. However,

since the co-occurrence of psychological mechanisms was not systematically varied, we were unable to disentangle their contributions independent of the historical context. Although beyond the scope of our paper, such an analysis may be possible using more sophisticated techniques. For example, in the domain of multi-alternative multi-attribute choice, Turner et al. (2018) use the switchboard technique, where each mechanism can be turned to different states (e.g., ON or OFF), to create compositional models, and thus systematically investigate the core psychological mechanisms underlying multi-attribute choice. Such techniques can potentially provide even better estimates of the contributions of different mechanisms in the domain of risky choice, and we hope that our work will inspire further research in this topic in the near future.

**Future work**

We have attempted a large-scale test of different risky decision models (and their corresponding psychological mechanisms) on a diverse set of experimental datasets. This approach is increasingly necessary given the growth of risky decision making research over the past few decades. Our model crowd approach also provides a promising way to model and analyze choice behavior by synthesizing the insights generated by dozens of models and allows us to quantitatively track the historical evolution of the field. The statistician George Box wittingly proclaimed that "all models are wrong but some of them are useful" (Box 1979). With the contribution crowd, we can evaluate models, either new or old, with regards to their unique contribution to the crowd, thus identifying which models are "useful" and to what degree.

Although the results of this analysis are largely robust across different datasets, and for different stimuli samples, in some cases the strength of certain models did depend on the design of gambles. An example is Experiment 2 of Pachur et al. (2018), whose design choice favored heuristic models, such as the better-than-average, minimax-regret and equiprobable heuristics.

These heuristic strategies do not involve the essential mechanisms of payoff and probability transformation and are thus unlikely to generalize to other settings (such as those involving randomly generated gambles). Future work can use our paradigm to better understand the effect of design choice on model behavior and model discrimination (see Navarro et al. 2004, Wagenmakers et al. 2004, Pitt and Myung 2009, Cavagnaro et al. 2013, Cavagnaro et al. 2016, He et al. 2020 for additional discussions). Additionally, although we have used an extremely large set of existing datasets to analyze model performance, all decision problems used in our analysis involve binary two-branch risky choices with full information. In the future, our approach could be extended to additional datasets, or types of problems. For example, researchers could examine how models perform in settings where more than two risky options are available (Venkatraman et al. 2014) or when the decisions are made under ambiguity (e.g. Ellsberg 1961).

We have addressed the theoretical and methodological challenges involved in modeling risky choice. These challenges are also common in other domains of decision research, such as intertemporal choice, decisions from experience, multi-attribute choice, social decision making, and strategic decision making (e.g. Frederick et al. 2002, Hertwig et al. 2004, Herzog and von Helversen 2018, Golman et al. 2020). In each of the domains there are a large number of competing behavioral models and pre-existing experimental datasets with considerable individual-level data. The large-scale model evaluation and model crowd approaches showcased in this paper can also be used to assess the state of the art in these areas and to further improve researchers' ability to predict people's choices. We look forward to future work that builds upon the numerous existing theories and extensive empirical data in decision science, in order to provide a cumulative, trans-disciplinary perspective on human choice behavior.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on the Information Theory*, Ed. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademia Kiado.

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.

Analytis, P. P., Barkoczi, D., & Herzog, S. M. (2018). Social learning strategies for matters of taste. *Nature Human Behaviour*, *2*(6), 415-424.

Armstrong, J. S. (2001). Combining forecasts. In *Principles of Forecasting* (pp. 417-439). Springer, Boston, MA.

Ashton, A. H., & Ashton, R. H. (1985). Aggregating subjective forecasts: Some empirical results. *Management Science*, *31*(12), 1499-1508.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science, 329*(5995), 1081-1085.

Becker, G. M., & McClintock, C. G. (1967). Value: Behavioral decision theory. *Annual Review of Psychology*, *18*(1), 239-286.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations Research*, *30*(5), 961-981.

Bell, D. E. (1985). Disappointment in decision making under uncertainty. *Operations Research*, *33*(1), 1-27.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *SiGKDD Explorations, 9*(2), 75-79.

Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, *1738*(5), 175-192. (Translated into English by L. Sommer in *Econometrica*, 1954, 22, 23-36.)

Bhatia, S., & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review, 124*(5), 678-687.

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, US: Lawrence Erlbaum Associates Publishers.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review*, *115*(2), 463-501.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, *127*(3), 1243-1285.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*(2), 409-432.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

Breiman, L. (1998). Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, *26*(3), 801-849.

Box, G. E. (1979). Robustness in the strategy of scientific model building. In *Robustness in Statistics* (pp. 201-236). Academic Press.

Budescu, D. V., & Chen, E. (2014). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267-280.

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.

Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, *5*(4), 325-370.

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: An optimized adaptive test. *Journal of Risk and Uncertainty, 52*(3), 233-254.

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

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Coombs, C. H., & Pruitt, D. G. (1960). Components of risk in decision making: Probability and variance preferences. *Journal of Experimental Psychology, 60*(5), 265-277.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise?. *Decision*, *1*(2), 79-101.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668-1674.

Delquié, P., & Cillo, A. (2006). Disappointment without prior expectation: a unifying perspective on decision under risk. *Journal of Risk and Uncertainty, 33*(3), 197-215.

Diecidue, E., & Van De Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, *49*(2), 683-700.

Dyer, J. S., & Jia, J. (1997). Relative risk—value models. *European Journal of Operational Research*, *103*(1), 170-185.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 51*(4), 380-417.

Edwards, W. (1955). The prediction of decisions among bets. *Journal of Experimental Psychology*, *50*(3), 201-204.

Edwards, W. (1956). Reward probability, amount, and information as determiners of sequential two-alternative decisions. *Journal of Experimental Psychology, 52*(3), 177-188.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, *12*(1), 473-498.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgement and choice. *Annual Review of Psychology*, *32*(1), 53-88.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*(1), 158-172.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*, 643-669.

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.

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.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*(2), 107-140.

Fiedler, S., & Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, *3*, 335.

Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. *The American Economic Review*, *67*(2), 116-126.

Frederick, S., Loewenstein, G., & O'donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351-401.

Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, *75*(7), 450-451.

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*, *110*(3), 605-639.

Galesic, M., Barkoczi, D., & Katsikopoulos, K. (2018). Smaller crowds outperform larger crowds and individuals in realistic task conditions. *Decision*, *5*(1), 1-15.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*(1), 1-58.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*(1), 107-143.

Goldstein, D. G., McAfee, R. P., & Suri, S. (2014, June). The wisdom of smaller, smarter crowds. In *Proceedings of the fifteenth ACM conference on Economics and computation* (pp. 471-488). ACM.

Golman, R., Bhatia, S. & Kane, P. (2020). The dual accumulator model of strategic deliberation and decision making. *Psychological Review, 127* (4), 477-504.

Handa, J. (1977). Risk, probabilities, and a new theory of cardinal utility. *Journal of Political Economy*, *85*(1), 97-122.

Hastie, R. (2001). Problems for judgment and decision making. *Annual Review of Psychology*, *52*(1), 653-683.

He, L., Zhao, W.J. & Bhatia, S. (2020). An ontology of decision models. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000231.

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.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517-523.

Herzog, S. M., & von Helversen, B. (2018). Strategy Selection Versus Strategy Blending: A Predictive Perspective on Single-and Multi-Strategy Accounts in Multiple-Cue Estimation. *Journal of Behavioral Decision Making*, *31*(2), 233-249.

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Performance, 21*(1), 40-46.

Hogarth, R. M., & Einhorn, H. J. (1990). Venture theory: A model of decision weights. *Management Science*, *36*(7), 780-803.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382-401.

Kahneman D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363-391.

Karmarkar, U. S. (1978). Subjectively weighted utility: A descriptive extension of the expected utility model. *Organizational Behavior and Human Performance*, *21*(1), 61-72.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge, England: Cambridge University Press.

Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. The Quarterly *Journal of Economics, 121*(4), 1133-1165.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience*, *13*(10), 1292-1298.

Lamberson, P. J., & Page, S. E. (2012). Optimal forecasting groups. *Management Science*, *58*(4), 805-810

Lattimore, P. K., Baker, J. R., & Witte, A. D. (1992). The influence of probability on risky choice: A parametric examination. *Journal of Economic Behavior & Organization*, *17*(3), 377-400.

Leland, J. W. (1994). Generalized similarity judgments: An alternative explanation for choice anomalies. *Journal of Risk and Uncertainty, 9*, 151–172.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124*(6), 762–794.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review, 125*(1), 1-32.

Loewenstein, G., O'Donoghue, T., & Bhatia, S. (2015). Modeling the interplay between affect and deliberation. *Decision, 2*(2), 55-81.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267-286.

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*(368), 805-824.

Loomes, G., & Sugden, R. (1986). Disappointment and dynamic consistency in choice under uncertainty. *The Review of Economic Studies*, *53*(2), 271-282.

Luan, S., Katsikopoulos, K. V., & Reimer, T. (2012). When does diversity trump ability (and vice versa) in group decision making? A simulation study. *PloS one*, *7*(2), e31043.

Machina, M. J. (1982). " Expected Utility" Analysis without the Independence Axiom. *Econometrica*, *50*, 277-323.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*(9), 987-996.

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276-299.

Marchiori, D., Di Guida, S., & Erev, I. (2015). Noisy retrieval models of over-and undersensitivity to rare events. *Decision*, *2*(2), 82-106.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77-91.

Mellers, B., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*(3), 332.

Mukherjee, K. (2010). A dual system model of preferences under risk. *Psychological Review, 117*(1), 243–255.

Müller-Trede, J., Choshen-Hillel, S., Barneron, M., & Yaniv, I. (2017). The wisdom of crowds in matters of taste. *Management Science*, *64*(4), 1779-1803.

Myung, J. I., & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116*(3), 499–518.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49*(1), 47-84.

Niculescu-Mizil, A., Perlich, C., Swirszcz, G., Sindhwani, V., Liu, Y., Melville, P., ... & Shang, W. X. (2009, December). Winning the KDD cup orange challenge with ensemble selection. In *KDD-Cup 2009 Competition* (pp. 23-34).

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*, 277-294.

Pachur, T., Mata, R., & Hertwig, R. (2017). Who dares, who errs? Disentangling cognitive and motivational roots of age differences in decisions under risk. *Psychological science*, *28*(4), 504-518.

Pachur, T., Suter, R. S., & Hertwig, R. (2017). How the twain can meet: Prospect theory and models of heuristics in risky choice. *Cognitive psychology*, *93*, 44-73.

Pachur, T., Schulte-Mecklenbeck, M., Murphy, R. O., & Hertwig, R. (2018). Prospect theory reflects selective allocation of attention. *Journal of Experimental Psychology: General*, *147*(2), 147-169.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General*, *143*(5), 2000-2019.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534-552.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1992). Behavioral decision research: A constructive processing perspective. *Annual Review of Psychology*, *43*(1), 87-131.

Pitz, G. F., & Sachs, N. J. (1984). Judgment and decision: Theory and application. *Annual Review of Psychology, 35*(1), 139-164.

Pitt, M. A., Kim, W., & Myung, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, *10*(1), 29-44.

Plonsky, O., Erev, I., Hazan, T., & Tennenholtz, M. (2017, February). Psychological forest: Predicting human behavior. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, *6*(3), 21-45.

Pollatsek, A., & Tversky, A. (1970). A theory of risk. *Journal of Mathematical Psychology*, *7*(3), 540-553.

Prelec, D. (1998). The probability weighting function. *Econometrica, 66*, 497-528.

Rapoport, A., & Wallsten, T. S. (1972). Individual decision behavior. *Annual Review of Psychology*, *23*(1), 131-176.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446-1465.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E.-J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review, 120*(1), 39-64.

Sheng, F., Ramakrishnan, A., Seok, D., Zhao, W. J., Thelaus, S., Cen, P., & Platt, M. L. (2020). Decomposing loss aversion from gaze allocation and pupil dilation. *Proceedings of the National Academy of Sciences*, *117*(21), 11356-11363.

Simonson, I., Carmon, Z., Dhar, R., Drolet, A., & Nowlis, S. M. (2001). Consumer research: In search of identity. *Annual Review of Psychology, 52*(1), 249-275.

Singmann, H., Kellen, D., Mizrak, E., & Öztekin, I. (2018). Using Ensembles of Cognitive Models to Answer Substantive Questions.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*(1), 1-39.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In *Computational Social Psychology* (pp. 311-331). Routledge.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature*, *38*(2), 332-382.

Stewart, N., Hermens, F., & Matthews, W. J. (2016). Eye movements in risky choice. *Journal of Behavioral Decision Making*, *29*(2-3), 116-136.

Stewart, N., Reimers, S., & Harris, A. J. (2015). On the origin of utility, weighting, and discounting functions: How they get their shapes and how to change their shapes. *Management Science*, *61*(3), 687-705.

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty, 32*(2), 101-130.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies, and nations. New York, NY: Doubleday.

Thorngate, W. (1980). Efficient decision heuristics. *Behavioral Science*, *25*(3), 219-225.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review, 125*(3), 329–362.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297-323.

Viscusi, W. K. (1989). Prospective reference theory: Toward an explanation of the paradoxes. *Journal of Risk and Uncertainty, 2*(3), 235-263.

Von Neumann, J., & Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton, NJ: Princeton University Press.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192-196.

Wagenmakers, E. J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*(1), 28-50.

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53-85.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review, 111*(2), 430-445.

Wu, G., Zhang, J., & Abdellaoui, M. (2005). Testing prospect theories using probability tradeoff consistency. *Journal of Risk and Uncertainty*, *30*(2), 107-131.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica: Journal of the Econometric Society*, 95-115.