In: R.S. St. Amant, D. Reitter, & E.W. Stacy (Eds.), Proceedings of the 23rd Annual Conference on Behavior Representation in Modeling & Simulation. Washington, DC: BRIMS Society, 2014.

Decisions from Experience: Could Models of Aggregate Choice Explain Individual Choices from Information Search?

Neha Sharma[1], *Varun Dutt*[1, 2]
[1]School of Computing and Electrical Engineering
[2]School of Humanities and Social Sciences
Indian Institute of Technology, Mandi, India
neha_sharma@students.iitmandi.ac.in, varun@iitmandi.ac.in

Keywords:
Aggregate choice, Individual choice, Experience, Sampling, Computational cognitive models, Error ratio

ABSTRACT: *Decisions from Experience (DFE) involve situations where decision makers search for information before making a final choice (e.g., looking-up washing machines from various companies before buying one). Conventionally, in DFE research, the final choice is aggregated over participants. However, this aggregation does not explain the individual participant choices. In this paper, we test the ability of top three DFE computational models of aggregate choice to account for choices at the individual level from information search. A Primed-Sampler (PS) model, a Natural-Mean Heuristic (NMH) model, and an Instance-Based Learning (IBL) model are calibrated to explain individual choices in the Technion Prediction Tournament (largest publically available DFE dataset).Our results reveal that all DFE models perform well to explain individual choices. Meta-analysis of the IBL Model (the model that explains the aggregate choice best) reveals that positive outcomes and high probabilities in risky options propel the model towards choosing risky options more often. We highlight the implications of our results for decisions made from information search.*

1. Introduction

More and more smart-phone companies are launching new mobile phones with similar designs and specifications almost every day. Generally, mobile stores allow people to touch, feel, and experience phones before a customer decides on buying a mobile-phone for real. Trying phones before buying them allows customers to search for information on how good the device is and what handset would suit them. The act of making choices based upon sampled information, however, is not restricted to selecting the best smart phones in the market. Rather, it is a common way of life when making a selection out of the given set of options (e.g., trying all delicacies in a sweet shop before making a final purchase, test driving cars before buying a particular one, collecting information from companies before choosing a career etc.). In fact, information search by sampling before a consequential choice constitutes an integral part of Decisions from Experience (DFE) research, where the focus is on explaining human decisions based upon one's experience with sampled information.

In order to study people's search and choice behaviors in the laboratory, a "sampling paradigm" has been proposed in the DFE research (Hertwig & Erev, 2009). In the sampling paradigm, people are presented with two or more options to choose between. These options are represented as blank buttons on a computer screen. People are first asked to sample as many outcomes as they wish from different button options (information search). Once people are satisfied with their sampling of the options, they decide from which option to make a single final choice for real. Human choice behavior in the sampling paradigm has so far been predicted at an aggregate level by the computational cognitive models. Here, the final choices made after sampling information in a problem is averaged across large number of participants and then this aggregate final choice is compared to a similar aggregate choice from computational cognitive models (Busemeyer &Wang, 2000; Gonzalez & Dutt, 2012; Lejarraga, Dutt, & Gonzalez, 2012). For example, the Primed-Sampler (PS) model, the Natural-Mean Heuristic (NMH) model, and the Instance-Based Learning (IBL) model are popular DFE algorithms for explaining aggregate choices (Erev et al., 2010; Gonzalez & Dutt, 2011). The PS model depends upon the recency of sampled information, where the model looks back a few samples on each option before making a final choice (Gonzalez & Dutt, 2011). On the other hand, the NMH model is a generic case of the PS model. In this model, one calculates the natural mean of outcomes observed on each option during sampling, making a final choice for the option with a larger natural mean (Hertwig,2011). Similarly, the IBL model (Gonzalez & Dutt, 2011) consists of experiences (called instances) stored in memory. Each instance's activation is a function of the frequency and recency of the corresponding outcomes observed during sampling in

different options. These activations are used to calculate the blended values for each option, where the option with the highest blended value is the one chosen at final choice. Prior DFE research has shown that, at the aggregate level, the PS, NMH, and IBL models exhibit superior performance compared to other computational models in the sampling paradigm (Erev et al., 2010; Gonzalez & Dutt, 2011). For evaluating these models at the aggregate level, a comparison has been made between a model's data and human data from the Technion Prediction Tournament (TPT) dataset (TPT being the largest publically known DFE dataset). However, up to now, none of the three DFE models have been evaluated in their ability to account for choice behavior at the individual participant level (i.e., in explaining the final choice of each human participant playing a problem in a dataset). If these models are able to account for choices at the aggregate level, then one expects that they might also be able to account for choices at the individual level. However, given that there are sources of noise in both the sampling data as well as in these models, it is likely that these models are no better than random chance in explaining choices at the individual level.

In this paper, our main goal is to evaluate how models, which explain choice behavior at the aggregate level, perform at the individual level. In order to evaluate models at the level of an individual participant, we use the largest publically available TPT dataset in the sampling paradigm (Erev et al., 2010). As described above, the sampling paradigm involves sampling outcome information available on options before making a final choice for real (Hertwig, 2011). In what follows, we detail the working of the three models described above. Then, we discuss the methodology of calibrating these models at the individual participant level. Next, we present the results of models' evaluation at the individual level. Finally, we close the paper by discussing the implications of our results for models of aggregate choice.

2. The Models

In this section, we detail the working of three popular DFE models that have been used to evaluate human choices at the aggregate level.

2.1 Prime Sampler (PS) Model

The PS model (Hertwig, 2011) employs a simple choice rule. In this model, participants are expected to take a sample of *k* draws from each option, and select the option with the highest sample mean. The exact value of *k* differs between observations (an observation is defined as a participant playing a problem in a dataset). The PS model assumes that the exact value of *k* for an observation is uniformly drawn as an integer between 1 and *N*, where *N* is a free parameter that is calibrated in the model. If the value of *k* for an observation is larger than the sample size of an option, then the *k* is restricted to that option's sample size. The final choice for each observation is determined by choosing the option with the highest sample mean based upon the last *k* draws. According to literature, the PS model has performed very accurately at predicting aggregated human choices in the sampling paradigm (Erev, Glozman, &Hertwig, 2008).

2.2 Natural Mean Heuristic (NMH) Model

The NMH model (Hertwig & Pleskac, 2010) involves the following steps:

Step 1. Calculate the natural mean of observed outcomes for each option by summing, separately for each option, all *n* experienced outcomes and then dividing by *n*.
Step 2. Choose the option with the largest natural mean (i.e., the option that had the best average outcome in the sampling phase).

Thus, the NMH model is a special case of the PS model, where *k* = an option's sample size. There are no free parameters in the NMH model. Like the PS model, this model has also performed very accurately at predicting aggregated human choices in the sampling paradigm (Hertwig & Pleskac, 2010).

2.3 Instance Based Learning (IBL) Model

The IBL model is based upon the ACT-R framework (Gonzalez & Dutt, 2011; 2012) and this model is known to predict human aggregate choices better than several DFE models, including those with assumptions similar to the PS and NMH models (Gonzalez & Dutt, 2011). In this model, every occurrence of an outcome on an option is stored in the form of an *instance* in memory. An instance is made up of the following structure: SDU, here S is the current situation (a number of blank option buttons on a computer screen), D is the decision made in the current situation (choice for one of the option buttons), and U is the goodness (utility) of the made decision (the outcome obtained upon making a choice). When a decision choice needs to be made, instances belonging to each option are retrieved from memory and blended together. The option that has the highest blended value is the one chosen. Blended values are a function of activation of instances being blended. Activation is a function of the frequency and recency of observed outcomes that occur on choosing options during sampling. In binary choice, the IBL model chooses one of two options by selecting the one with the highest blended value (Gonzalez & Dutt, 2011; 2012). The blended value of option *j* (e.g., a gamble that pays $4 with .8 probability or $0) is defined as

$$V_j = \sum_{i=1}^{n} p_i x_i \qquad \cdots (1)$$

where x_i is the value of the U part of an instance *i* (e.g., either \$4 or \$0, in the previous example) and p_i is the probability of retrieval of that instance from memory (Lebiere, 1999). Because x_i is the values of the U part of an instance *i*, the number of terms in the summation changes when new outcomes are observed within an option *j* (and new instances corresponding to observed outcomes are created in memory). Thus, *n*=1 if *j* is a safe option with one possible outcome. If *j* is a risky option with two possible outcomes, then *n*=1 when one of the outcomes has been observed on an option (i.e., one instance is created in memory) and *n*=2 when both outcomes have been observed (i.e., two instances are created in memory). Given that the choice rule simply entails selecting the option with the highest blended value, the model is naturally suited for problems that involve choices among more than two options (Lejarraga, Dutt, & Gonzalez, 2012). At any trial *t*, the probability of retrieval of an instance *i* is a function of the activation of that instance relative to the activation of all instances created within that option, given by

$$P_{i,t} = \frac{e^{A_{i,t}/\tau}}{\sum_j e^{A_{j,t}/\tau}} \qquad \cdots (2)$$

Where, τ is random noise defined as $= \sigma . \sqrt{2}$ and σ is a free noise parameter.

Noise in Equation (2) captures the imprecision of recalling past experiences from memory. The activation of an instance corresponding to an observed outcome in a given trial is a function of the frequency of the outcome's past occurrences and the recency of the outcome's past occurrences. At each trial *t*, activation A of an instance *i* is

$$A_{i,t} = \sigma \ln\left(\frac{1-\gamma_{i,t}}{\gamma_{i,t}}\right) + \ln \sum_{t_i \epsilon \{1,\cdots,t-1\}} (t-t_i)^{-d} \qquad \cdots (3)$$

where *d* is a free decay parameter; $\gamma_{i,t}$ is a random draw from a uniform distribution bounded between 0 and 1; and t_i is each of the previous trial indexes in which the outcome corresponding to instance *i* was observed. The IBL model has two free parameters that need to be calibrated: *d* and σ. The *d* parameter controls the reliance on recent or distant sampled information. Thus, when *d* is large (> 1.0), then the model gives more weight to recently observed outcomes in computing instance activations compared to when *d* is small (< 1.0). The σ parameter helps to account for the sample-to-sample variability in an instance's activation.

2.4 The Coin Toss (CT) Model

The CT model is used as a baseline model and it represents chance performance. In this model, we compare the value of a random number between [0, 1] with probability = 0.5. In a binary-choice task, if the random number value < 0.5, then the model chooses the final choice as one option; otherwise, the model chooses a final choice as the other option. When simulated, for a binary-choice task, this model is expected to produce close to 50% accuracy in explaining participants' individual choices. As the probability is fixed at 0.5, this model contains no free parameters.

3. Method

3.1 The Technion Prediction Tournament

The Technion Prediction Tournament (TPT) (Erev et al., 2010) was a competition in which several participants were subjected to an experimental setup, the "e-sampling condition." In this condition, participants sampled the two blank button options in a binary-choice problem before making a final choice for one of the options. During sampling, participants were free to click both button options one-by-one and observe the resulting outcome. Participants were asked to press the "choice stage" key when they felt that they had sampled enough (but not before sampling at least once from each option). The outcome of each sample was determined by the structure of the relevant problem. One option corresponded to a safe choice: Each sample provided a medium (M) outcome. The other option corresponded to the payoff distribution of a risky choice: Each sample provided a High (H) payoff with some probability (pH) or a low (L) payoff with the complementary probability (1 - pH). At the choice stage, participants were asked to select once between the two options. Their choice yielded a random draw of one outcome from the selected option and this outcome was considered at the end of the experiment to determine the final payoff.

Competing models submitted to TPT were evaluated following the generalization criterion method (Busemeyer &Wang, 2000), by which models were fitted to choices made by participants in 60 problems (the estimation set) and later tested in a new set of 60 problems (the test set) with the parameters obtained in the estimation set. The 120 problems consisted of choice between a safe option and a risky option as described above. The M, H, pH, and L in a problem were generated randomly, and a selection algorithm was used so that the 60 problems in each set differed in its M, H, pH, and L from other problems. Among the 60 problems in each set, 20 problems were such that the probability range for pH was less than or

equal to 0.1 (low probability); 20 problems were such that the probability range for pH was between 0.1 and 0.9 (medium probability); and, 20 problems were such that the probability range for pH was greater than or equal to 0.9 (high probability). Similarly, among the 60 problems in each set, 20 problems were such that all three outcomes, H, L, and M outcomes were positive (positive domain); 20 problems were such that one of the outcomes, H, L, and M, was negative and others were positive (mixed domain); and 20 problems were such that all three outcomes, H, L, and M, were negative (negative domain). In the models and human data described here, we have considered an individual model/human participant playing a problem in a dataset (competition or estimation) as an "observation." Also, all model parameters have been calibrated by using the entire TPT dataset that consisted of 120 problems and 2,370 observations. For more details about the TPT, please refer to Erev et al. (2010).

3.2 Model Executions

Models submitted to the TPT were evaluated only according to their ability to account for aggregate risk-taking behavior (i.e., the proportion of risky choices aggregated across participants and problems (Erev et al., 2010). In this paper, we account for this risk-taking at the individual participant level. For this purpose, the sampling choices from human data are fed to a model. Then, after sampling, a final choice (risky or safe) made by a model observation is evaluated against a final choice made by a corresponding human. In order to compare human and model final choices for each observation, we evaluate an "error ratio" (i.e., the ratio of incorrectly classified final choices between model and human observations divided by the total number of observations). Firstly, for each observation in human data, we determine the final choice (risky or safe). A similar final choice is then derived for a model observation and this derived choice is compared to the choice made by the corresponding human observation. The final choices from each of the three models are compared to 2,370 human observations, i.e., the total number of participant-problems-set combinations in TPT dataset. For a model, the error-ratio is calculated as:

Error Ratio = (RS+SR)/ (RS+SR+RR+SS)

Where, RS was the number of observations where the model predicted a safe choice but the human made a risky choice. SR was the number of observations where the model predicted a risky choice but human made a safe choice. Similarly, the RR and SS were the number of observations, where a model predicted the same choice (risky or safe) as made by a human observation in human data. The smaller the value of the error ratio, the more accurate is a model in accounting for individual choices of human participants. For some observations, a model was equally likely to choose a safe or a risky choice. Such cases were discarded from the error ratio calculation and were termed as uncategorized (UN) cases. Thus, more are the number of uncategorized cases; the poorer is the corresponding model's algorithm in accounting for the size of human data.

In the PS model, an integer value was drawn between 1 and N. For the purposes of model calibration, a maximum value of N was assumed to be 216. This choice of maximum value was justifiable as 216 is the maximum sample size in the TPT dataset (Erev et al.,2010). For each new value of N from 1 to 216, the PS model was run 5 times over the set of 2,370 observations (i.e., for each run, 2,370 observations were used in the model). Error ratio was computed for each of the 5 runs and these five 5 ratios were then averaged for calculating the average error ratio. The N value for which the average error ratio was minimized was taken as the calibrated N value. The choice of 5 runs of the model is because it enabled us to account for the randomness present in the model.

The NMH model did not possess any free parameters. The model's calibration involved only the natural mean computation for each observation based on the outcomes observed on both the risky and safe options during sampling. In NMH model, the final choice was made for an option that had the highest natural mean. This model was run across 2,370 observations only once since it did not possess any noise or random computations.

The IBL model described here has two free parameters d and that were calibrated using a genetic algorithm program. The genetic algorithm repeatedly modified a population of individual parameter tuples in order to find the tuple that minimized the error ratio. In each generation, the genetic algorithm selected individual parameter tuples randomly from a population to become parents and used these parents to select children for the next generation. Over successive generations, the population evolved toward an optimal solution. The population size used here was a set of 20 randomly-selected parameter tuples in a generation (each parameter tuple was a particular value of d and σ). The mutation and crossover fractions were set at 0.1 and 0.8, respectively, for an optimization over 150 generations in Matlab. For each parameter tuple, the IBL model was run 5 times across 2,370 observations. Across the 5 runs, the model's average error ratio was computed by averaging the error ratios from each run. The parameter tuple that minimized the average error ratio across 150 generations were reported as the calibrated parameters for the IBL model.

Finally, the CT model was also run across 2,370 observations for 5 times and the average error ratio was

calculated by averaging the error ratios obtained across the 5 runs.

4. Results

In this section, we present the results obtained from the three different models described above.

In the PS model, we varied N as an increasing integer between 1 and 216 and for each value of N we ran the model 5 times across 2,370 observations. For each N value, the average error ratio was calculated across the 5 model runs. The best average error ratio across 5 runs was found to be 0.173. This error ratio occurred at N=17. Figure 1 shows the average error ratio results obtained from the PS model for different values of N. As shown in the figure, for the first few values of N, the error ratio reduced rapidly as the size of N increases (up to N = 17 samples). However, the error-ratio value saturated to a smaller proportion after increasing N further. Thus, there was not much variation in the error ratio from N=18 to N=216.

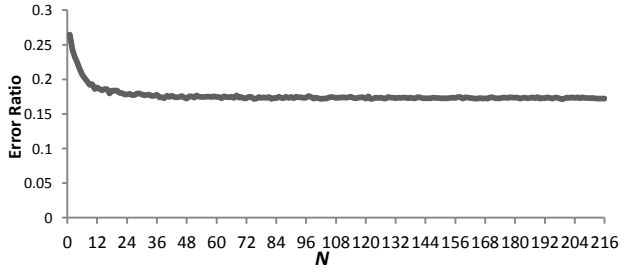

Figure 1: Results from the PS model. The value of parameter *N* was varied between 1 and 216. The best average error ratio = 0.173 for *N*=17. There were 33 UN observations on average across the 5 runs.

Table 1 shows the individual-level results from the PS model for the best value of *N* = 17. As shown in Table 1, the SS combinations constituted 42.9% of total combinations. This SS proportion was the highest amongst other combinations (RR, RS, SR, and UN). The RR combinations had the second highest value of 38.5%. The combined average of both the combinations formed about 82% of correctly predicted choices by the model. In contrast, the erroneous RS and SR ratios were at 7.67% and 9.40% respectively. The uncategorized (UN) observations were close to 33 out of 2,370 observations. Thus, there were about 1.4% observations that could not be explained by this model. The PS model's best average error ratio is almost 32% better than the average error ratio of 50% resulting from the CT model. Thus, the PS model explained a large proportion of human dataset fairly accurately.

Table 1.1 Results from the calibrated PS model. The error ratio = 0.173 for N = 17. There were 33 UN observations across the dataset.

Combinations from Human Data and Model	Number of Observations	Percentage of 2370 Observations
SS	1019	42.9
RR	913	38.5
RS	182	7.67
SR	223	9.40
UN	33	1.39

Next, we investigated the performance of the NMH model and Table 2 shows the results from this model. The model's average error ratio equaled 0.161 across 2,370 observations. In the NMH model, the SS and RR values accounted for 43.6% and 39.4% of all 2,370 observations, respectively. The best value of error ratio was for the SS combinations, where about 43.6% of the 2,370 human observations were accounted by the model. The second best value accounted by the model was that for the RR combinations. In contrast, the erroneous RS and SR combinations from the model were only 6.8% and 9.1% of the 2,370 human observations, respectively. The error ratio obtained from the NMH model is only marginally lower than that obtained from the PS model. In fact, to predict an observation's final choice, the NMH model took into account the outcomes across all samples in an observation while the PS model considered only the last few samples before the final choice. However, the NMH model, like the PS model, could not classify 26 observations (the UN observations). Thus, like the PS model, the NMH model was able to explain a large dataset. The model showed an improvement of 34% over the coin toss model and performed efficiently at the individual level.

Table 1.2. The results from the NMH model. The average error ratio = 0.161. There were 26 UN observations across the dataset.

Combinations from Human Data and Model	Average Number of Observations	Percentage of 2370 Observations
SS	1033	43.6
RR	934	39.4
RS	162	6.8
SR	215	9.1
UN	26	1.09

Next, we evaluated the IBL model's ability to account for individual final choices in the TPT dataset. The results from IBL model are presented in Table 3. The best calibrated values of *d* and σ in the IBL model were found to be 9.42 and 0.32, respectively. The large *d* value exhibited reliance on recency during sampling. Also, the small σ value exhibited lesser sample-to-sample variability in instance activations. The calibrated IBL model produced 38.8% of SS combinations and 35.6% of RR combinations, respectively. Although the percentages of SS and RR combinations were similar, they both were slightly less than those found in the NMH and PS models, respectively. Having lower values for the SS and RR combinations reduces the accuracy of the IBL model compared to the NMH and PS models. In contrast, the erroneous SR and RS combinations were 11.1% and 14.4 % respectively from the IBL model and both these percentages were slightly greater than those obtained in the NMH and PS models. Thus, the human safe choices were predicted as risky by the model for about 11.1% of total observations. This erroneous classification is about 2% higher than the same erroneous classification from the NMH model. Moreover, the RS combinations have a 7.57% higher error compared to the NMH model. Based upon above statistics, the IBL model's performance was similar but slightly inferior to the PS and NMH models. However, the number of uncategorized (UN) cases resulting from the IBL model was zero. Thus, the IBL model's algorithm was able to predict a choice for all 2,370 human observations. In the PS and NMH models, the UN observations accounted for almost 1%-2% of the total observations. Since the uncategorized cases decrease the efficiency of the model, the IBL model had an advantage over the other two models on the UN cases. Furthermore, as the average error ratio from the IBL model was 0.255, the model shows 25% superior performance compared to the CT model.

Table1.3. Results from the calibrated IBL model. The calibrated value of parameters d=9.42 and σ=0.32. The average error ratio= 0.255. There was no UN observation across the 5 runs.

Choice Combinations from Human Data and Model	Average Number of Observations across 5 Runs	Percentage of 2370 Observations
SS	920.8	38.8
RR	844.8	35.6
RS	262.2	11.1
SR	342.2	14.4
UN	0	0

In order to develop a deeper understanding about the results from the IBL model, we analyzed the model behavior in different outcome domains and probability ranges in the TPT datasets. Table 4 shows how the IBL model results compared to human data in the three domains (negative, mixed, and positive). As seen in Table 4, the model chooses risky when the average difference in blended values between the risky and safe options is positive and the model chooses safe when this average difference in blended values is negative (the average BV Difference is the difference between the BV of risky option – the BV of safe option, averaged over all observations in the domain). In general, the number of observations for which the model makes a risky choice and the human makes a safe choice (i.e., the SR cases) increases from the negative domain towards the positive domain (78 observations in the negative domain, 120 observations in the mixed domain, and 138 observations in the positive domain). This increase in SR observations across the domains is explained by positive differences in blended values from the IBL model. Similarly, the number of observations where the model makes a safe choice and the human makes a risky choice (i.e., the RS cases) are fewer for the mixed and positive domains compared the negative domain (109 observations in the negative domain; whereas, only 67 and 79 observations in the mixed and positive domains). These RS observations are also explained by the negative blended value difference from the model. In general, the error ratio from the IBL model remained similar across the three domains with a slightly higher ratio of 0.27 in the positive domain. This latter increase is due to the increase in the SR cases and a decrease in the SS cases.

Table1.4 Results of meta-analysis from the IBL model across the three domains

Domain	Choice Combinations from Human Data and Model	Average BV Difference(BV Risky – BV Safe)	Average Number of Observations	Error Ratio
Negative	SS	-2.12	343	0.24
	SR	1.50	78	
	RS	-1.58	109	
	RR	3.32	260	
Mixed	SS	-2.51	307	0.24
	SR	1.28	120	
	RS	-1.77	67	
	RR	2.22	296	
Positive	SS	-2.87	277	0.27
	SR	1.88	138	
	RS	-1.69	79	
	RR	2.41	296	

Table 5 shows meta-analysis results for the three probability ranges for pH (low, medium and high). As per our analysis above, if the difference in blended value of risky option and safe option was positive, then the IBL model would choose the risky option; otherwise, the IBL model would choose the safe option. The SR cases have 14 observations in the low probability range, 130 observations in the medium probability range, and 192 observations in the high probability range. Here, the average blended value difference is positive and hence the model makes incorrect risky predictions. On the other hand, the RS cases have 137 observations in the low probability range, 90 observations in the medium probability range, and 28 observations in the high probability range. The blended value is negative for these RS cases and therefore the model predicts a safe final choice, where the human choice is risky. In general, the error ratio increased from the low probability range to the high probability range and this increase was attributed to the rapid increase in the SR cases and a rapid decrease in the SS cases.

Table1.5 Results of meta-analysis from the IBL model across the three Probability Ranges.

Probability	Choice Combinations from Human Data and Model	Average BV Difference (BV Risky – BV Safe)	Average Number of Observations	Error Ratio
Low	SS	-0.82	548	0.19
	SR	7.37	14	
	RS	-0.78	137	
	RR	7.77	92	
Medium	SS	-4.52	245	0.28
	SR	2.43	130	
	RS	-2.51	90	
	RR	3.88	323	
High	SS	-5.47	134	0.28
	SR	0.57	192	
	RS	-3.57	28	
	RR	0.60	437	

The Table 6 shows the average error ratios from the PS, NMH, and IBL models. The PS model gave an error ratio of 0.173. The NMH model gave an error ratio of 0.161, which was very close to the best value of error ratio from the PS model. Similarly, the IBL model gave an error ratio of 0.255. The NMH model considers the complete sample size of each observation as opposed to the PS model, where the PS model takes into consideration only last few samples of each observation. In this regard, we can conclude that the PS model is more efficient than the NMH model as it uses a much smaller proportion of samples for about the same error ratio compared to the NMH model. Both the PS and NHM models have a significant edge over the CT model with 32% to 34% smaller error ratios. However, both the PS and NMH models also have 26 and 33 UN cases which decrease their efficiency. The IBL model has an error ratio of 0.255 which is higher than the other two models; however, the IBL model does not have any UN observations.

Table 1.6. Summary of results from the three DFE models

Model	Parameters	UN Observations	Error ratio
PS	N = 17	26	0.173
NMH	-	33	0.161
IBL	d=9.42,σ=0.32	00	0.255

6. Discussion and Conclusion

Till recently, literature in judgment and decision making had compared cognitive models of choice by evaluating their performance at the aggregate level (Gonzalez & Dutt, 2011; 2012). In these comparisons, the average risk-taking at final choice from the model has been compared to the average risk-taking at final choice from human data. However, in this paper, we evaluate a model's performance at the individual participant level. Thus, we attempt to match predictions from a model to those from the human data for each observation, where the observation's sampling is fed to the model.

Three popular and competing models of aggregate human choice were evaluated on their abilities towards explaining individual human choices. Overall, our results reveal that all the three models of aggregate choice performed well at the individual level (error ratio <= 25%). However, some of the models (e.g., NMH and PS) were not be able to account for the entire human dataset because of equal expected values for the two options in these models, rendering an impossible prediction. Although the IBL model does not perform as well as the other two models, the IBL model uses blending and the distinctness of the blended values across the two options yields a choice prediction for each human observation in the dataset. Thus, the IBL model's strength is in its ability to account for human data across a large number of observations.

Our results also allowed us to compare the ability of the three models to explain individual safe and risky choices observed in human data. In this regard, it was found that all three models find it easier to explain individual safe choices compared to individual risky choices (there were

greater proportion of SS combinations compared to RR combinations across all three models). This observation could be due to the presence of a single medium outcome in the safe option compared to the presence of both the low and high outcomes in the risky option. Because of the single outcome in the safe option its expected value or blended value might not fluctuate as rapidly as that under the risky option with two very different outcomes.

In addition, all three models reported a higher proportion of erroneous SR combinations compared to the erroneous RS combinations. Thus, all three models predicted a risky final choice for human observations, where humans actually end-up making a safe choice. This finding could also be due to the fact that the risky option contained two outcomes compared to the single outcome in the safe option. Thus, the greater variability experienced in the risky option drove a model to making more risky choices; whereas, the same variability drove the human to make safe choices.

Furthermore, the IBL model's meta-analysis revealed that both the three domains and the three probability ranges influenced the number of mismatched SR or RS observations from human data and the model. In general, the direction of these mismatched results from the IBL model could be explained by the direction of the differences between the blended values of the risky and safe options. A deeper analysis of how the blended values change across samples in these mismatched observations will likely open-up new possibilities for future research.

We believe that this paper augurs a beginning of a larger research program that plans to launch an in-depth investigation of the presence of RS and SR cases among influential models of experiential choice. As part of future research, we would like to extend our investigation to problems where both options are risky rather than current problems where one option was safe and the other option was risky. By testing models in problems with two risky options, we would be able to investigate the role of variability in affecting contradictory human and model choices as depicted by the SR and RS cases. Furthermore, we plan to calibrate models on the TPT dataset and test them on a different dataset, where the problem structure is different from that of the problems in the TPT. Furthermore, in this paper, we took top three competing models of experiential choice; however, as part of future research, we plan to extend this investigation to a larger set of DFE models that cover other theoretical ideas.

7. Acknowledgements

The authors are grateful to Dr. Ido Erev, Technion - Israel Institute of Technology, for providing the Technion Prediction Tournament dataset. Also, the authors would like to thank Indian Institute of Technology, Mandi for providing computational resources for this project.

8. References

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

Erev, I., Glozman, I., & Hertwig, R. (2008). What impacts the impact of rare events. *Journal of Risk Uncertainty* 36:153–177.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., & Hau, R.(2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Mak*ing, 23, 15–47.

Gonzalez, C.,& Dutt, V. (2012). Refuting data aggregation arguments and how the instance-based learning model stands criticism: A reply to Hills and Hertwig. *Psychological Review, Vol* 119(4),893-898.

Gonzalez, C., & Dutt, V. (2011). Instance-Based Learning: Integrating Sampling and Repeated Decisions From Experience. *Psychological Review*,118(4), 523-551.

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

Hertwig, R. (2012). The psychology and rationality of decisions from experience. *Synthese*, 187, 269-292.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples*?*. *Cognition*, 115, 225–237.

Lebiere, C. (1999). Blending: An ACT–R mechanism for Aggregate retrievals. *Paper presented at the 6th Annual ACT–R Workshop at George Mason University*.Fairfax County, VA.

Lejarraga.Dutt.Gonzalez.(2012).Instance-Based Learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*,25: 143–153.

Author Biographies

NEHA SHARMA is a Ph.D. scholar in School of Computing and Electrical Engineering at the Indian Institute of Technology Mandi, India. Her current research interests include Decisions from Experience and Intelligent Systems.

VARUN DUTT is Assistant Professor, School of Computing and Electrical Engineering and School of Humanities and Social Sciences, Indian Institute of Technology, Mandi, India. He is also the Knowledge Editor of the English daily, *Financial Chronicle*. His current research interests are in dynamic decision making, environmental decision making, and modeling human behavior.