Modeling Choices in Delay Discounting

Dirk U. Wulff

Max Planck Institute for Human Development, Berlin, Germany

Wouter van den Bos

Max Planck Institute for Human Development, Berlin, Germany

Word count: 1,619

Corresponding author:

Dirk U. Wulff

Max Planck Institute for Human Development

Center for Adaptive Rationality

Lentzeallee 94, 14195 Berlin, Germany

Dirk.wulff@gmail.com

Acknowledgments:

We thank Ralph Hertwig, David Kellen, Corinna Laube, and Robert Lorenz for helpful comments, and Susannah Goss for editing.

Modeling Choices in Delay Discounting

In a recent article, Ericson and colleagues (Ericson, White, Laibson, & Cohen, 2015) compared traditional utility-discounting models with a set of heuristic models of intertemporal choice. Traditional utility-discounting models assume that the greater the delay in receiving an option, the more its value is discounted, whereas heuristic models of intertemporal choice assume that decisions are based on simple rules involving attribute-wise comparisons. Consistent with earlier reports (Dai & Busemeyer, 2014), Ericson and colleagues concluded from their cross-validation approach that heuristic models (specifically the novel *intertemporal choice heuristic* or ITCH model) explain intertemporal choices better than discounting models do. More surprisingly, their results showed that all discounting models performed nearly at chance level, and did not outperform even the baseline model (Figure 1A). If these findings were valid, they would have major implications for hundreds of studies using discounting models. However, here we demonstrate that these conclusions are premature. Models of both classes are in fact rather good at predicting choice, and the jury is still out on which (type of) model is best.

Three aspects of the modeling approach used impacted the conclusions of Ericson et al. (2015). First, they used varying auxiliary assumptions, making it difficult to identify the source of differences in model performance (see Blavatskyy & Pogrebna, 2010). Second, the data used for model fitting and predicting were drawn from the entire pool of data aggregated across participants. However, this approach is valid only if one assumes that all participants relied on the same decision mechanism (see Estes, 1956; Estes & Maddox, 2005). Third, Ericson et al. chose mean absolute deviation to evaluate the models. This choice of loss function is, however, inappropriate when dealing with probabilistic models of choice. That is, mean absolute deviation does not select models that best predict the underlying choice probabilities (Buja, Stuetzle, and Shen 2005), although this is key in light of the widely accepted assumption of probabilistic choice (Rieskamp, 2008, see Supplementary Material). Moreover, by fitting models using maximum likelihood and evaluating them using mean absolute deviation, Ericson et al. relied on non-matching criteria for fitting and evaluation, which can heavily bias model evaluation (Elliott, Ghanem, & Krüger, 2016; Gneiting, 2011).

We re-analyzed the Ericson et al. (2015) data to see how the models fared when assessed under different, possibly more appropriate, assumptions. First, we gauged the impact of implementing varying auxiliary assumptions[1]. Here, we focused on two specific model adjustments: (i) removing the bias parameter in the heuristic models and (ii) implementing a different choice rule for the discounted utility models (Luce, 1959/2005; Pleskac, 2015). Second, we extended the initial analyses by comparing the models under cross-validation of subject-level data. Third, we evaluated the models under different evaluation criteria. Here, we report mean absolute deviation (MAD) as used by Ericson et al. as well as average negative log-likelihood (logLoss), which matches the maximum likelihood criterion used for fitting the models. Finally, we implemented the well-established dual-parameter hyperbolic model (HYPER2; Green & Myerson, 2004) and a more recent neuroscience-inspired double exponential model (SYSTEM; van den Bos & McClure, 2013). Full specifications of models, auxiliary assumptions, and evaluation criteria, as well as supporting discussion and analyses, are provided in the Supplementary Material.

[1] Auxiliary assumptions enable theories, e.g., the ITCH, to make empirical predictions. Whether an assumption is auxiliary or not depends on whether it is explicitly included in the theory of interest or not (see Lakatos, 1970). For example, in the present case, choice models must thus be considered auxiliary as they are not an element of the to be tested models, e.g., the ITCH and the DRIFT. However, when an investigation is (also) about the validity of choice models (e.g., Stott, 2006), then choice models are no longer auxiliary.

Beyond replicating the results of Ericson et al. (2015; see Figure 1A), our analyses revealed three critical insights. (i) *Auxiliary assumptions matter*: Removing the bias parameter severely impacted the performance of the heuristic models (under MAD loss function; Figure 1A and 1B). In addition, implementing the power choice rule boosted the performance of all discounting models. As a result, some adjusted discounting models performed on par with the best performing heuristic models for both aggregate (Figure 1A) and subject-level data (Figure 1B). (ii) *Beware of aggregation*: For subject-level data, predictive power was much better than at the aggregate level for all models. (iii) *Evaluation criteria matter*: Using logLoss instead of MAD reversed the pattern of results: All of the adjusted discounting models outperformed the heuristic models for subject-level data (similar patterns of results were obtained using mean squared error and zero-one loss; see supplementary material). Moreover, using logLoss reversed the impact of the bias parameter. Under this evaluation criterion, lacking a bias parameter actually improved the performance of two of the heuristic models.

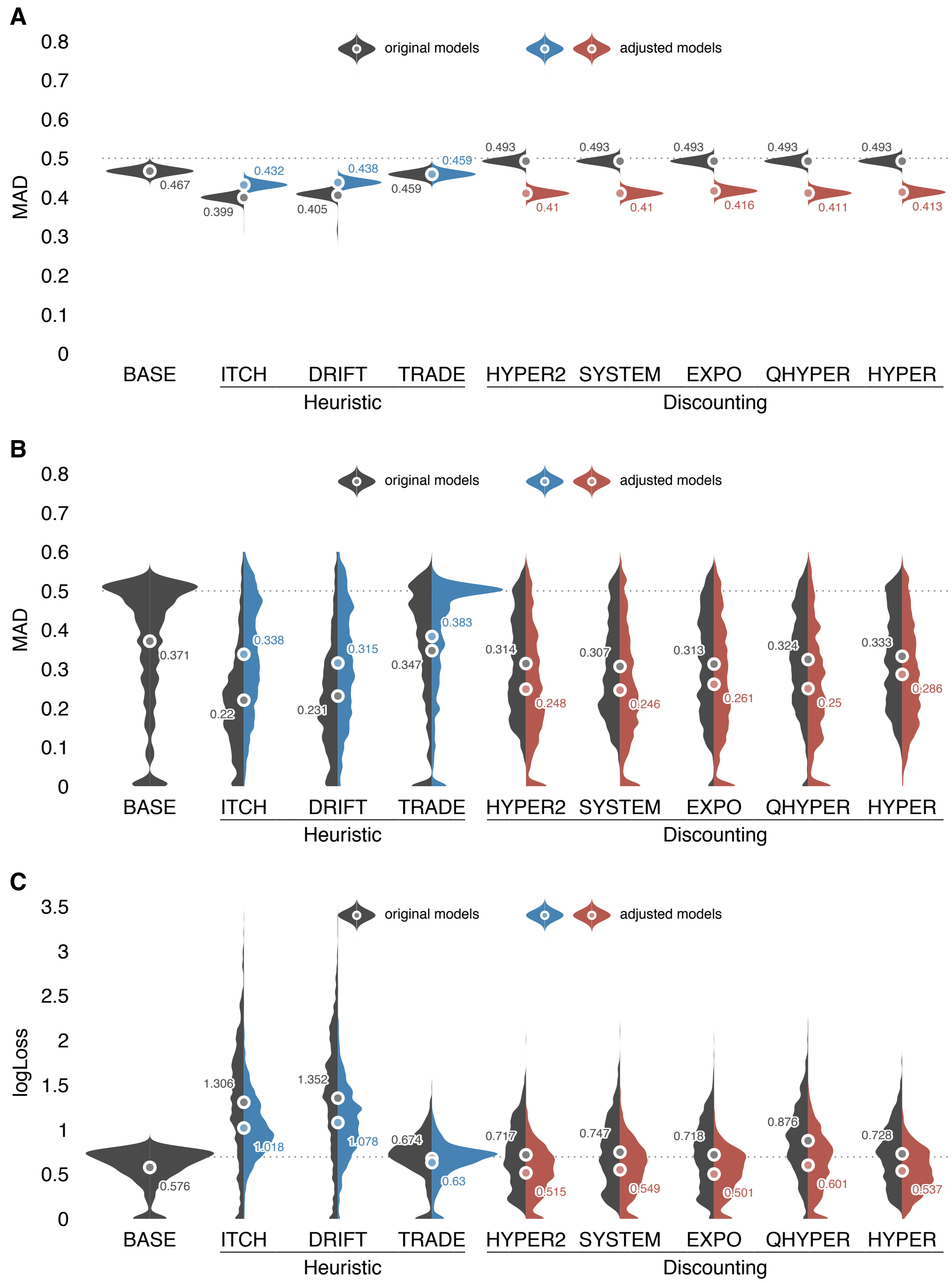

Figure 1. Violin plots showing the distribution of models' prediction performances across 10,000 cross-validation repetitions based on (A) aggregate and (B & C) subject-level data, for both the original and the adjusted model specifications. The dark gray distributions are those presented in the original article. The heuristic models were adjusted by removing the bias parameter (blue), and the discounting models were adjusted by changing the choice rule from an exponential to a power rule (red). Figures A and B show results under the MAD loss function for comparison with the original article; Figure C shows results under the logLoss loss function. Circles represent the mean of the respective distributions. The dashed lines correspond to a coin-flip model predicting .5 for every choice. The coin-

flip model predicts chance performances of MAD = .5 and logLoss = –log(.5) = .693. Results are shown collapsed over the five conditions reported in Ericsson et al., but were stable across conditions.

Our reanalysis of the Ericson et al. (2015) data demonstrated first that, contrary to the findings of Ericson et al., the discounting models were much better than chance at predicting choice, and accounted for the data at least as well as the heuristic models. Second, we demonstrated the importance of auxiliary assumptions. Specifically, the bias parameter, which could be regarded as a pre-evaluative process, such as the application of a strict aspiration level of receiving something now (see Stewart, Reimers, & Harris, 2014; Wulff, Hills, & Hertwig, 2015), appeared to lend the heuristic models substantial flexibility that improved performance under MAD, but negatively impacted performance under logLoss (see Myung, 2000, for a discussion of model flexibility). The use of a power choice rule also affected model performance substantially by boosting the predictive power of the discounting models. Note that the goal of this comment is not to recommend specific auxiliary assumptions, but rather to highlight the importance of exploring them. For instance, we are aware that the power choice rule cannot accomodate negative outcomes and is thus unsuited to explain behavior across a wider set of problems. Also, our exploration of auxiliary assumptions is not nearly exhaustive, and it is quite possible that adding others may again alter the results. However, such an outcome would only underline our conclusion. Third, evaluating the models on the individual level resulted in a dramatic improvement in the performance of the heuristic models and the discounting models. This result strongly suggests heterogeneity in the decision-making processes of individuals (see Marewski & Schooler, 2011, and Rieskamp & Otto, 2006). Fourth, the results of the model comparisons depended heavily on the choice of loss function. Specifically, when logLoss was used instead of MAD, the pattern of results reversed such that the adjusted discounting models outperformed the heuristic models. Choices among loss functions are not arbitrary; there are strong theoretical reasons to choose a loss function that suits the data to be predicted (in this case, the probability of a choosing the larger later option; see Merkle & Steyvers, 2013), and that matches the loss function used to fit the models (Elliott, Ghanem, & Krüger, 2016; Gneiting, 2011). In the present case, this means that logLoss is the most appropriate choice of loss function, and that most weight should be placed on the results associated with that choice of loss function (see Supplementary Material for a more extensive argument). In sum, our re-analyses result in three clear recommendations: (1) explore not only different core theories but also the auxiliary assumptions, (2) use subject-level data, and (3) select the same loss function for training and testing in cross-validation.

Another (fourth) issue, but one we could not address in our re-analyses, is the selection of choice problems. For successful model comparisons, the design must generate data that can distinguish between models (Donkin, Newell, Kalish, Dunn, & Nosofsky, 2015; Navarro, Pitt, & Myung, 2004; Wulff & Pachur, 2016). This may, however, not be the case for the present study design. For example, there is an imbalance in the decision problems implemented in Ericson et al. (2015); the maximum outcome was $101,000, i.e., roughly four times the yearly per capita income in the USA, whereas maximum waiting time was only six weeks. It is known that the range of stimuli used can severely affect model recovery (Broomell & Bathia, 2014). Moreover, it is possible that short maximum delays explain the surprisingly good performance of the exponential model (EXPO), a model often shown to be unfeasible (e.g. Mazur, 1987).

Finally, although our analyses showed that discounting models may provide a useful quantitative measure of choice behavior, the models may fail as descriptions of the underlying cognitive processes (van den Bos & McClure, 2013). The heuristic models, in contrast, do suggest plausible cognitive mechanisms, and there are good reasons to believe that people may rely on attribute-wise comparisons in decision-making (Su et al., 2013). For instance, a substantial subset of participants was best fit by one of the heuristic models (see Supplementary Material). Our findings raise the question whether selecting models based on choice patterns is sufficient to make strong claims about the underlying cognitive processes. One fruitful avenue to further corroborate such claims is to use process data, such as eye-tracking (Johnson, Schulte-Mecklenbeck, & Willemsen, 2008) or neuroimaging data (Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016), to further constrain the model space.

References

Blavatskyy, P. R., & Pogrebna, G. (2010). Models of stochastic choice and decision theories: Why both are important for analyzing decisions. *Journal of Applied Econometrics, 25*(6), 963–986. http://dx.doi.org/10.1002/jae.1116

Broomell, S. B., & Bhatia, S. (2014). Parameter recovery for decision modeling using choice data. *Decision, 1*(4), 252–274. http://dx.doi.org/10.1037/dec0000020

Buja, A., Stuetzle, W., and Shen, Y. (2005), "Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications," manuscript, available at www-stat.wharton.upenn.edu/~buja/.

Dai, J., & Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General, 143*(4), 1489–1514. http://dx.doi.org/10.1037/a0035976

Donkin, C., Newell, B. R., Kalish, M., Dunn, J. C., & Nosofsky, R. M. (2015). Identifying strategy use in category learning tasks: A case for more diagnostic data and models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(4), 933–?. http://dx.doi.org/10.1037/xlm0000083

Elliott, G., Ghanem, D., & Krüger, F. (2016). Forecasting conditional probabilities of binary outcomes under misspecification. *Review of Economics and Statistics*. Advance online publication.

Ericson, K. M. M., White, J. M., Laibson, D., & Cohen, J. D. (2015). Money earlier or later? Simple heuristics explain intertemporal choices better than delay discounting does. *Psychological Science, 26*(6), 826–833. http://dx.doi.org/10.1177/0956797615572232

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134–140. http://dx.doi.org/10.1037/h0045156

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12*(3), 403–408. http://dx.doi.org/10.3758/BF03193784

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, *106*(494), 746-762. http://dx.doi.org/10.1198/jasa.2011.r10138

Green, L., & Myerson, J. (2004). A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin, 130*(5), 769–792. http://dx.doi.org/10.1037/0033-2909.130.5.769

Lakatos, I. (1970). Falsification and the methodology of scientific research. In Lakatos, I. & Musgrave, A.(Eds.) Criticism and the growth of knowledge (pp. 170-196). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/cbo9781139171434.009

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Mineola, NY: Dover Publications. (Original work published 1959)

Johnson, E. J., Schulte-Mecklenbeck, M., & Willemsen, M. C. (2008). Process models deserve process data: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*(1), 263–273. http://dx.doi.org/10.1037/0033-295X.115.1.263

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*(3), 393–437. http://dx.doi.org/10.1037/a0024143

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A. Nevin, & H. Rachlin (Eds.), *Quantitative* *analyses of behavior: Vol. 5. The effect of delay and of intervening events on reinforcement value* (pp. 55–73). Hillsdale, NJ: Erlbaum.

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis, 10*(4), 292–304. http://dx.doi.org/10.1287/deca.2013.0280

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44*(1), 190–204. http://dx.doi.org/10.1006/jmps.1999.1283

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. http://dx.doi.org/10.1016/j.cogpsych.2003.11.001

Pleskac, T.J., 2015. Decision and choice: luce's choice axiom. In: J. D. Wright (Ed.), *International encyclopedia of the social & behavioral sciences* (2nd ed., pp. 895-900). Oxford, UK: Elsevier. http://dx.doi.org/10.1016/b978-0-08-097086-8.43031-x

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*(2), 207–236. http://dx.doi.org/10.1037/0096-3445.135.2.207.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6), 1446-1465. http://dx.doi.org/10.1037/a0013646

Stewart, N., Reimers, S., & Harris, A. J. (2014). 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. http://dx.doi.org/10.1287/mnsc.2013.1853

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. Journal of Risk and uncertainty, 32(2), 101-130. http://dx.doi.org/10.1007/s11166-006-8289-6

Su, Y., Rao, L. L., Sun, H. Y., Du, X. L., Li, X., & Li, S. (2013). Is making a risky choice based on a weighting and adding process? An eye-tracking investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1765–1780. http://dx.doi.org/10.1037/a0032861

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2015). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *NeuroImage, 128*, 96–115. http://dx.doi.org/10.1016/j.neuroimage.2015.12.030 .

van den Bos, W., & McClure, S. M. (2013). Towards a general model of temporal discounting. *Journal of the Experimental Analysis of Behavior, 99*(1), 58–73. http://dx.doi.org/10.1002/jeab.6

Wulff, D. U., Hills, T. T., & Hertwig, R. (2015). How short-and long-run aspirations impact search and choice in decisions from experience. *Cognition, 144*, 29–37. http://dx.doi.org/10.1016/j.cognition.2015.07.006 .

Wulff, D. U., & Pachur, T. (2016). Modeling valuations from experience: A comment on Ashby and Rakow (2014). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 158–166. http://dx.doi.org/10.1037/xlm0000165