Supplementary Materials for: The Role of Passing Time in Decision-Making

Nathan J. Evans[ab], Guy E. Hawkins[b] and Scott D. Brown[b]

[a] Department of Psychology, University of Amsterdam, The Netherlands

[b] School of Psychology, University of Newcastle, Australia

Analysis with "simple" fixed thresholds diffusion model

In the main text we compared a "full" fixed thresholds diffusion model to a "simple" collapsing thresholds diffusion model – which we believe is both the most theoretically interesting and practical comparison for the reasons outlined in the main text. The "full" fixed thresholds model includes between-trial variability in drift rate, starting point, and non-decision time, while the "simple" collapsing thresholds model did not contain between-trial variability in any parameters. However, in the cases where the collapsing thresholds model was found to be superior, our comparison leaves two potential explanations for why the collapsing thresholds model provided the better DIC value: either 1) the collapsing thresholds were able to capture certain aspects of the data, or 2) the best model was actually a fixed thresholds "simple" diffusion model, and the fixed thresholds model lost in the complexity-corrected DIC calculation because of the additional flexibility provided by the between-trial variability parameters that was not actually required to adequately account for the data.

In order to address this potential limitation, we also tested a "simple" variant of the fixed thresholds diffusion model to the data of each experiment, in the same manner as the models compared in the main text. The comparison of all three models (collapsing thresholds, "full" fixed thresholds, "simple" fixed thresholds) on DIC weights for each experiment can be seen in Figure S1. For Experiment 2 (the deadline experiment), no participant had any appreciable weight in favor of the simple fixed thresholds diffusion model, meaning that its inclusion had no impact on the results. For Experiment 1 (reward rate emphasis), most participants had no weight in favor of the simple fixed thresholds diffusion model, and the minority of participants that had non-zero weight did not change the trends, again meaning that its inclusion again had very little impact on the results. However, for Experiment 3 (speed emphasis), a considerable number of participants showed

a substantial weight in favor of the simple fixed thresholds diffusion model, meaning that the assessment of the fixed thresholds diffusion model in the main text may have been disadvantaged for some participants due to the inclusion of the between-trial variability parameters. To check whether this was the case, we compared the collapsing thresholds model to the best (i.e., lowest DIC) fixed thresholds model for each participant in each experiment, shown in Figure S2. However, none of the findings appear to be qualitatively different from those reported in the main text, and the pattern of results remains identical.

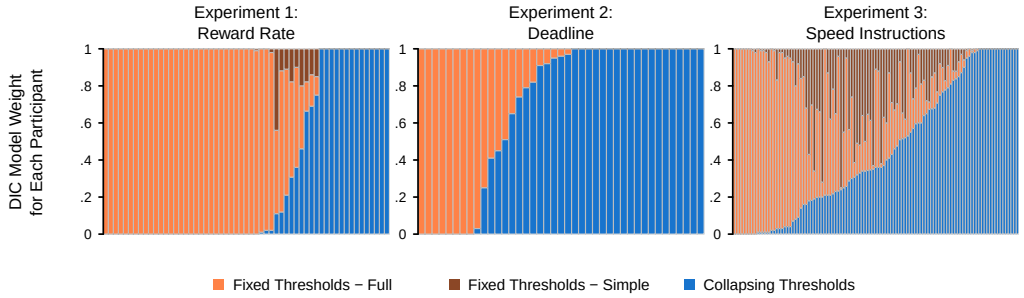

Figure S1. Results from Experiments 1-3 (columns) for the DIC analysis including the "simple" fixed thresholds diffusion model. The DIC weights were calculated as described in the method section of Experiment 1 in the main text, with the x-axis showing different participants (ordered by their weight in favor of the collapsing thresholds models), and y-axis the DIC weight associated with each model.

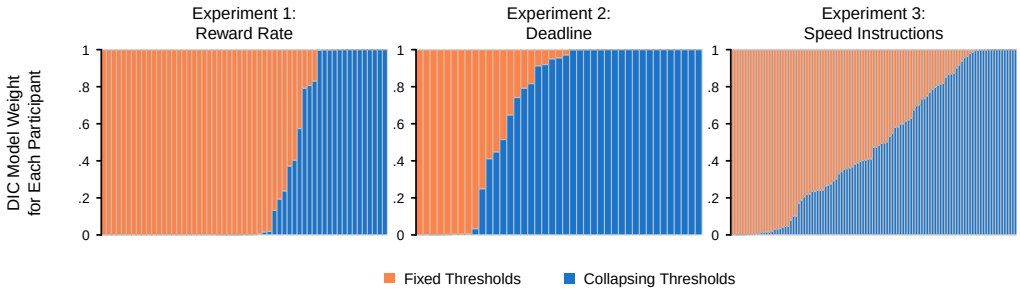

Figure S2. Results from Experiments 1-3 (columns) for the DIC analysis, with the fixed thresholds diffusion model for each participant represented by the simple *or* full diffusion variant, whichever provided the better DIC value. The DIC weights were calculated as described in the method section of Experiment 1, with the x-axis showing different participants (ordered by their weight in favor of the two models), and y-axis the DIC weight associated with each model.

Analysis without contamination process

In the main text we compared fixed and collapsing thresholds models that both contained a "contamination" process, where the response time distribution was assumed to be made up of x proportion random responses, and $1 - x$ proportion responses based on the model, where x was a free parameter. Importantly, this contamination process allowed for the possibility of participants making responses in some cases that were not the result of the standard decision-making process. Although contamination processes are commonly thought to be an auxiliary assumption that should not have a major impact upon inferences, we found that failing to include a contamination process led to different results, particularly in Experiment 3 (speed emphasis). The comparison between the collapsing and fixed thresholds models described in the main text – though without the contamination process – can be seen in Figure S3. Overall, the results appear to have shifted toward the fixed thresholds model relative to the results reported in the main text that incorporated a contamination process, which for Experiment 3 results in some overall preference in favor of the fixed thresholds model.

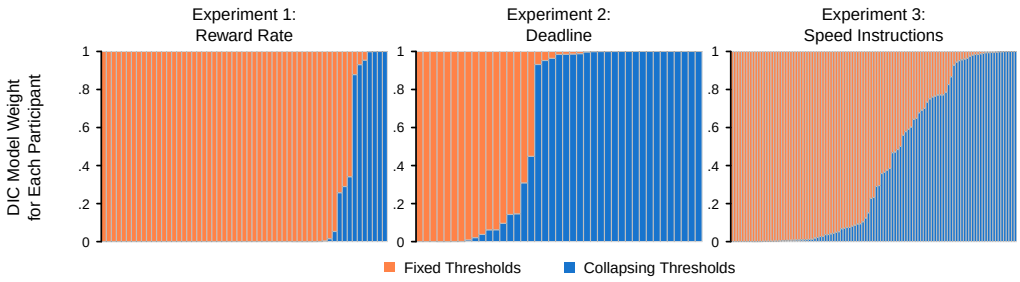

Figure S3. Results from Experiments 1-3 (columns) for the DIC analysis, with the fixed and collapsing thresholds diffusion models for each participant represented by the non-contamination variants. The DIC weights were calculated as described in the method section of Experiment 1, with the x-axis showing different participants (ordered by their weight in favor of the two models), and y-axis the weight associated with each model.

To determine whether our inferences should be based upon the contamination or non-contamination versions of the models, we compared all four models (fixed and collapsing thresholds, with and without the contamination process) on DIC weights for each experiment. As can be seen in Figure S4, most participants appear to be best accounted for by the contamination models, though many participants have large weights in favor of a non-contamination model. However, the general pattern of results appear to suggest that in cases where a non-contamination model is preferred, the equivalent contamination model has the next-best DIC, meaning that the overall pattern of results is maintained even when including the non-contamination models in the comparison. Figure S5 attempts to present this trend more clearly, comparing the best fixed thresholds model (i.e., contamination or non-contamination, whichever had the better DIC) and the best collapsing thresholds model (again, contamination or non-contamination, whichever had the better DIC) for each participant. The results from this analysis are qualitatively similar to those in the main text, suggesting that although an explicit assumption of no contamination

in any model changes the overall pattern of results, including models with and without a contamination process into the model comparison maintains the same pattern of results.

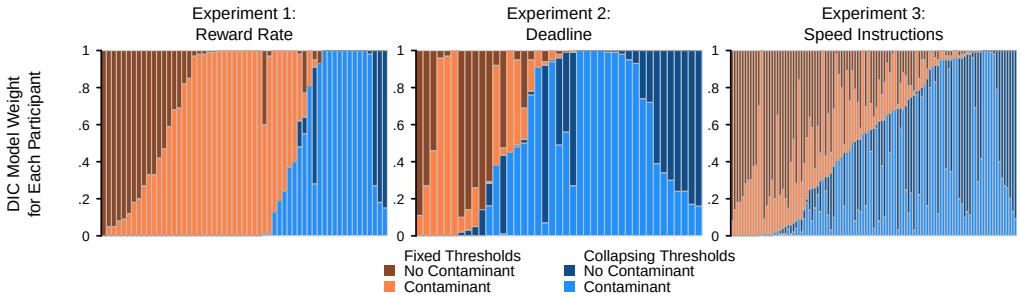

Figure S4. Results from Experiments 1-3 (columns) for the DIC analysis including the fixed and collapsing thresholds diffusion models with and without a contamination process. The DIC weights were calculated as described in the method section of Experiment 1, with the x-axis showing different participants (ordered by their overall weight in favor of both types of models), and y-axis the weight associated with each model.

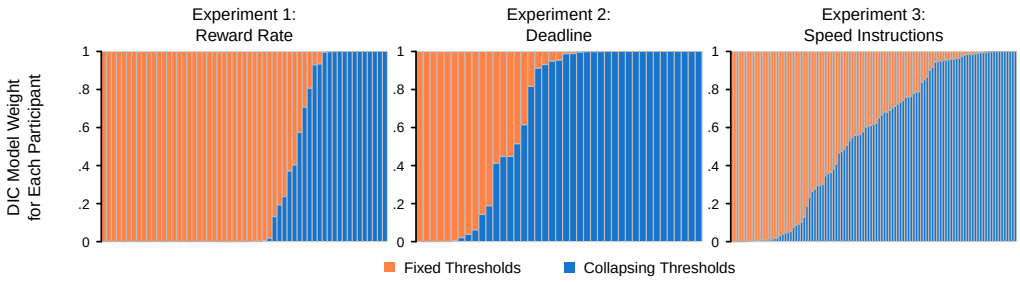

Figure S5. Results from Experiments 1-3 (columns) for the DIC analysis, with the fixed and collapsing thresholds diffusion models for each participant represented by either a contamination or non-contamination variant, whichever provided the better DIC value. The DIC weights were calculated as described in the method section of Experiment 1, with the x-axis showing different participants (ordered by their weight in favor of the two models), and y-axis the weight associated with each model.

Analysis of excluded participants

As mentioned in the main text, we defined our participant exclusion criteria for each experiment before analyzing the data. In addition, we based these exclusion criteria on the performance that we believed participants should be able to achieve if they 1) were engaged with the task, 2) understood how to complete the task, and 3) followed the instructions. However, our exclusion criteria did result in a large number of participants being excluded from Experiments 2 and 3, and different researchers may have different opinions on what should constitute "inadequate" performance from participants. Therefore, we also fit the data from the excluded participants in each experiment with the fixed thresholds and collapsing thresholds models described in the main text, and compared these models based on the DIC weights. However, for computational efficiency, we fit these subjects individually using Bayesian estimation, rather than the Bayesian hierarchical estimation used in the main text. We also included the performance of each participant under each potential exclusion criteria variable, so that readers can assess whether or not they believe each participant performed the task adequately (and therefore, whether the results of a participant should be considered legitimate). However, note that one participant from Experiment 2 could not be included, as *all* of their trials resulted in misses.

The analysis of these additional participants can be seen in Figure S6, and how each participant performed relative to the exclusion criteria can be seen in Tables S1, S2, and S3. For Experiment 1, where only a very small minority were excluded, 4 participants showed strong evidence for collapsing thresholds and 2 showed strong evidence for fixed thresholds. Although this pattern differs from the included participants (i.e., most participants displayed strong evidence in favor of fixed thresholds in the main text), these few participants are not enough to change the overall trend seen in these data, suggesting that the inclusion of these participants would not have changed the overall pattern of results

from the main text. For Experiment 2, 16 participants showed strong evidence for collapsing thresholds, 1 showed some evidence for collapsing thresholds, and 12 showed strong evidence for fixed thresholds. This result is consistent with the included participants in the main text, where most participants showed strong evidence in favor of collapsing thresholds, suggesting that the inclusion of these participants would not have changed the overall pattern of results. For Experiment 3, 31 participants showed strong evidence for collapsing thresholds, 1 showed some evidence for collapsing thresholds, 2 showed fairly ambiguous evidence, 2 showed some evidence for fixed thresholds, and 11 showed strong evidence for fixed thresholds. In general, these participants appear to show much more decisive evidence than those in the main text (i.e., most participants displayed fairly ambiguous evidence in the main text), and the inclusion of these participants would shift the overall trend in the data toward collapsing thresholds.

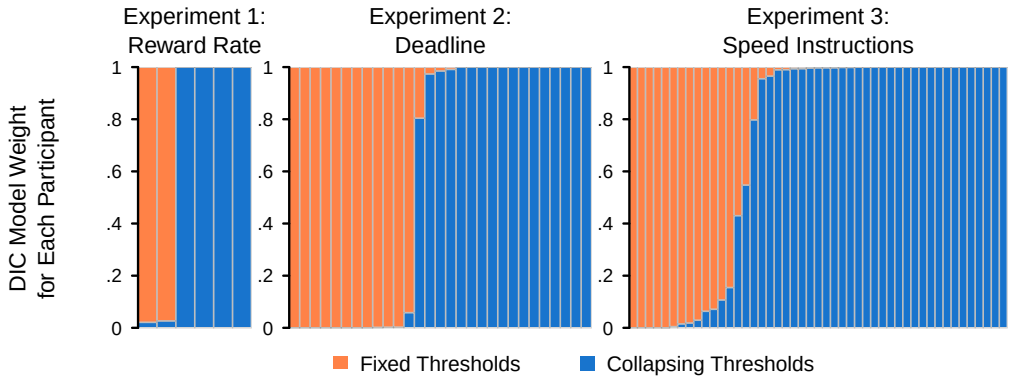

Figure S6. Results from Experiments 1-3 (columns) for the DIC analysis for the participants that were excluded from the primary analyses reported in the main text. The DIC weights were calculated as described in the method section of Experiment 1, with the x-axis showing different participants (ordered by their weight in favor of the two models), and y-axis being the weight associated with each model.

Table S1: Excluded participants from Experiment 1, with their DIC weight in favor of the collapsing thresholds model, and their performance on each of the relevant experiment exclusion criteria.

DIC weight (COLL)	ACC (overall)
1	0.55
0.03	0.54
1	0.58
1	0.56
0.02	0.51
1	0.56

Table S2: Excluded participants from Experiment 2, with their DIC weight in favor of the collapsing thresholds model, and their performance on each of the relevant experiment exclusion criteria.

DIC weight (COLL)	ACC (easy)	MISS
1	0.85	0.08
1	0.74	0.01
0	0.89	0.03
0.99	0.98	0.36
0	0.5	0.85
1	0.89	0.02
0.98	0.87	0.03
1	0.56	0.06
0	0.56	0.04
1	0.75	0.04
0	0.5	0.02
0.06	0.84	0.1
0	0.56	0.01
1	0.55	0.06
1	0.58	0.04
0	0.74	0.11
0.8	0.67	0.52
0	0.8	0.03
0	0.86	0.01
0	0.66	0.22
1	0.83	0.04
1	0.89	0.05
1	0.85	0.45
1	0.53	0.03
1	0.52	0.02
0	0.9	0
0	0.89	0.04
0.97	0.79	0.01
1	0.38	0.21

Table S3: Excluded participants from Experiment 3, with their DIC weight in favor of the collapsing thresholds model, and their performance on each of the relevant experiment exclusion criteria.

DIC weight (COLL)	ACC (easy)	MRT
1	0.6	1.67
0.03	0.92	1.61
1	1	1.69
1	1	1.99
1	0.98	1.64
1	0.81	1.59
1	0.96	1.7
1	0.65	2.17
1	1	1.59
0.96	0.83	2.09
1	0.98	1.68
1	0.97	2.37
0.97	0.98	1.71
0.43	0.98	1.57
1	0.69	1.23
1	1	1.74
1	0.96	1.61
0.55	0.81	1.49
0.01	0.64	3.09
1	1	1.67
0.99	0.58	1.47
0	0.54	0.66
0.06	0.42	0.55
1	0.88	1.68
1	0.83	1.14
1	0.52	1.45
0.02	0.85	0.79
0	0.51	0.83
0.15	0.85	1.38
0.8	0.98	1.59
0	0.9	0.67
0.11	0.85	1.12
0	0.85	1.41
1	0.98	2.14
1	0.94	1.65
1	0.75	1.05
1	0.88	1.44
0	0.89	0.73
0	0.9	0.75
1	1	1.69
1	0.98	1.99
0.99	1	1.7
0.99	0.94	1.65
1	1	1.67
0.07	0.85	1.07
1	0.98	1.7
0.99	0.98	1.63

Model Selection with DIC using Individual vs. Group Parameters

The comparison of the fixed and collapsing thresholds models within our paper focused on individual-level DIC values, which we used to assess how consistent the findings were across participants, where a large number of participants strongly favoring one model was considered strong evidence in favor of that model. Readers may wonder why we didn't instead calculate "group-level" values for the DIC metric, where a single DIC metric is calculated for each model over all participants in the experiment. These values would provide a single comparison for the entire data set, providing a clear answer to which model is superior, and have been commonly used in the past for AIC/BIC/DIC/WAIC (e.g., Rae, Heathcote, Donkin, Averell, & Brown, 2014; Evans, Brown, Mewhort, & Heathcote, 2018). Our reason for calculating individual-level DIC over group-level DIC was twofold. Firstly, previous studies in the debates about fixed vs. collapsing bound models have focused on individual-level metrics (Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Palestro, Weichart, Sederberg, & Turner, 2018), meaning that our study remains consistent with the types of inferences made in those studies. Secondly, after calculating the group-level DIC values, closer inspection of these results suggested that the conclusions of group-level DIC values were misleading in some cases (note that the same criticism would apply to group-level WAIC, or any other group-level metric for that matter). The problem is that the group-level analyses, which are based on some averages over different individual-level statistics, were representative of a small minority of participants who showed overwhelming evidence in favor of one model.

The group-level DIC values are shown as difference scores at the top of each panel in Figure S7, where positive values support the collapsing thresholds model. They indicate that the collapsing thresholds model is clearly the best model in all data sets, which would lead one to conclude that people adopt collapsing thresholds across the three different ex-

perimental settings. This is in clear contrast to the results reported in the main text, where most participants in Experiment 1 showed strong evidence in favor of fixed thresholds. The reason for this conflict can be seen in each panel of Figure S7, which plots the estimated contamination probability in the collapsing thresholds model (x-axis) and the DIC value in favor of the collapsing thresholds model (i.e., the difference in DIC between the fixed and collapsing thresholds models; y-axis). There appears to be a very small number of participants (5) in Experiment 1 that have extremely strong DIC values in favor of the collapsing thresholds model, which overwhelm the much larger number of participants who have strong, but not as strong, evidence in favor of fixed thresholds. Interestingly, the participants with overwhelming evidence in favor of collapsing thresholds also have relatively large contamination probabilities estimated – much larger than is typical in the literature. This suggests that these participants may have also given a large portion of unreliable responses. Therefore, we believe that the individual-level DIC analysis provides a more accurate, complete picture of the overall preference across participants for each model, and that group-level model selection values can sometimes be deceptive due to their sensitivity to extremely strong outliers.

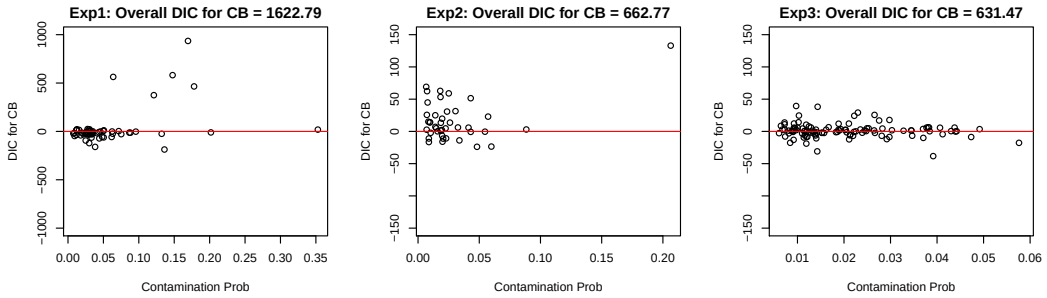

Figure S7. Results from Experiments 1-3 (columns) which show each participants' estimated contamination probability under the collapsing thresholds model (x-axis) and the DIC difference values (y-axis), where positive values indicate evidence in favor of collapsing thresholds and negative values indicate evidence in favor of fixed thresholds. The top of each panel provides the group-level DIC value, which in all cases is in favor of the collapsing thresholds model.

Stability of performance over blocks

For the main analyses, we excluded the first 21 blocks of trials from Experiment 1, which we based on previous studies (Evans & Brown, 2017; Evans, Bennett, & Brown, 2018). In those studies, participants took approximately 15 blocks to stabilize on a single strategy when given reward rate feedback. For Experiments 2 and 3, we only excluded the first block of trials to allow participants to be sufficiently practiced at the task, as they did not receive any mid-experiment feedback on their performance. However, readers may be interested in whether there appeared to be any change in performance across the blocks of trials included in the analysis, as this may reflect a change in strategy over blocks.

To assess whether performance appeared to change over blocks, we used default Bayesian ANOVAs (Rouder, Morey, Speckman, & Province, 2012) in the program JASP (JASP Team, 2018) to check for any strong evidence of changes in mean response time of correct responses or accuracy over blocks. For Experiment 1, there was strong evidence *against* a change in either mean response time ($BF_{01} = 66.7$) or accuracy ($BF_{01} = 1241.3$) over the included blocks of trials (22-30). This suggests that performance was stable across the blocks that we analyzed in Experiment 1. For Experiment 2, there was strong evidence *against* a change in accuracy ($BF_{01} = 166.7$) over the included blocks of trials (2-10), though some evidence *for* a change in mean response time ($BF_{10} = 6.1$). This suggests that performance was fairly stable within the blocks that we analyzed for Experiment 2, as there was no strong evidence in favor of a change in either variable. For Experiment 3, there was strong evidence *against* a change in accuracy ($BF_{01} = 10.1$) over the included blocks of trials (2-5), though strong evidence *for* a change in mean response time ($BF_{10} = 73780.3$). This suggests that performance may have changed over the blocks that we analyzed for this experiment. However, excluding additional blocks from Experiment 3 would have resulted in the total number of trials assessed becoming very small for response time models (e.g.,

Lerche, Voss, & Nagler, 2017), meaning that such results may be unreliable. However, readers may wish to take the results of Experiment 3 with a grain of salt, as there may have been a change in strategy over blocks, and this may explain some of the ambiguity found for the results in the main text.

Reward rate optimality analysis

In order to assess whether participants who adopted collapsing thresholds in Experiment 1 seemed to be doing so to maximize reward rate, we compared their achieved reward rate with the best possible reward rate achievable under a single fixed threshold. We used a similar method to Evans and Brown (2017) to find the best possible reward rate, where all parameters for each person – apart from the decision threshold – were fixed at the values that minimized the deviance (i.e., the values used for the pD calculation in the DIC metric). We then performed a grid search over threshold values between 0.01 and 4 in increments of 0.01, in order to find the threshold value that maximized the reward rate. Reward rate was calculated by:

$$\frac{PC}{MRT + ITI + FDT + (1 - PC) \times ET}$$

where MRT is mean response time, PC is the probability of a correct response, ITI is the inter-trial interval (100ms in our experiment), FDT is the feedback display time (300ms in our experiment), and ET is the error timeout (500ms in our experiment). The reward rate that would have been obtained from each possible threshold was calculated by simulating 10,000 trials from each experimental condition – using the method of Evans (2019) – and taking the reward rate as above, with the threshold that resulted in the highest reward rate being the "optimal" fixed threshold. Table S4 displays the reward rates for each participant, and the best possible reward rate that they could achieve using a fixed threshold. In general, almost every participant (55/57) shows a lower reward rate than the best possible reward rate under a fixed threshold, and the few participants who show a higher reward rate only do so to a minor extent, which may be attributed to noise in the optimality calculation process. Therefore, this suggests that participants who adopted collapsing thresholds did not do so because it moved them above the best possible fixed

threshold, and may explain why so few participants adopted collapsing thresholds.

Table S4: Reward rate, and optimal reward rate under a fixed threshold, Experiment 1 participants

Actual RR	"Optimal" RR	DIC weight (COLL)
0.48	0.52	0.00
0.55	0.61	1.00
0.41	0.42	1.00
0.54	0.58	0.00
0.45	0.51	1.00
0.32	0.39	1.00
0.43	0.47	0.00
0.45	0.46	1.00
0.68	0.72	0.00
0.43	0.44	0.00
0.55	0.59	1.00
0.51	0.61	0.00
0.54	0.54	0.00
0.55	0.60	0.00
0.43	0.52	0.00
0.53	0.58	0.00
0.35	0.41	1.00
0.55	0.56	0.00
0.44	0.55	0.01
0.52	0.61	1.00
0.42	0.51	0.24
0.55	0.59	0.37
0.48	0.57	0.00
0.55	0.57	0.00
0.47	0.51	0.00
0.43	0.57	1.00
0.53	0.61	0.00
0.51	0.55	0.00
0.53	0.65	0.40
0.59	0.63	0.00
0.55	0.62	0.00
0.62	0.69	0.13
0.56	0.67	1.00
0.57	0.60	0.00
0.49	0.53	0.00
0.48	0.56	0.19
0.45	0.47	0.02
0.50	0.61	0.57
0.48	0.53	0.00
0.71	0.74	1.00
0.55	0.60	0.00
0.56	0.58	0.00
0.31	0.35	1.00
0.46	0.52	0.00
0.61	0.69	0.02
0.36	0.40	1.00
0.52	0.58	0.00
0.42	0.55	0.00
0.35	0.38	0.89
0.50	0.53	0.00
0.45	0.45	0.00
0.49	0.56	0.00
0.44	0.46	0.00
0.43	0.53	1.00
0.46	0.49	0.00
0.43	0.52	0.81
0.44	0.53	0.81

References

Evans, N. J. (2019). A method, framework, and tutorial for efficiently simulating models of decision-making. *Behavior Research Methods*.

Evans, N. J., Bennett, A. J., & Brown, S. D. (2018). Optimal or not; depends on the task. *Psychonomic Bulletin & Review*, 1–8.

Evans, N. J., & Brown, S. D. (2017). People adopt optimal policies in simple decision-making, after practice and guidance. *Psychonomic Bulletin & Review*, *24*(2), 597–606.

Evans, N. J., Brown, S. D., Mewhort, D. J., & Heathcote, A. (2018). Refining the law of practice. *Psychological Review*, *125*(4), 592.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*(6), 2476–2484.

JASP Team. (2018). *JASP (Version 0.9)[Computer software].* Retrieved from `https://jasp-stats.org/`

Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods*, *49*(2), 513–537.

Palestro, J. J., Weichart, E., Sederberg, P. B., & Turner, B. M. (2018). Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychonomic Bulletin & Review*, 1–24.

Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The hare and the tortoise: Emphasizing speed can change the evidence used to make decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1226.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374.