Two-stage task with increased state space complexity to assess online planning

Jungsun Yoo
Department of Cognitive Science
University of California, Irvine
Irvine, CA 92697
jungsun.yoo@uci.edu

Aaron M. Bornstein
Department of Cognitive Science
University of California, Irvine
Irvine, CA 92697
aaron.bornstein@uci.edu

Abstract

Humans use an internal model to predict and navigate through a series of decisions in reinforcement learning (RL) tasks, referred to as planning. Studies show that planning predicts performance in such tasks, but they only provide a partial description because they do not separate local deliberation (i.e., online planning) from using plans made ahead of time (i.e., offline planning). To address this gap, we introduce a variant of the canonical two-stage task (TST), called the *multinomial* TST, that discourages planning beforehand by increasing state-space complexity (SSC). Here, we report behavioral results from three versions of multinomial TST with increasing SSC, in addition to the canonical TST (total $N = 418$). Consistent with the hypothesis that increasing SSC would lead to an increase in online planning, we found that increasing SSC induced longer response time (RT) during first-stage (but not second-stage) choices. We next decomposed RT into separable components of pre-trial and on-demand evaluation by fitting a novel variant of a reinforcement learning diffusion decision model (RLDDM). Model fits revealed that first-stage drift rate and starting point both showed influence of model-based values, but that, as SSC increased, this influence was stronger in drift rate and weaker in starting point. Further, we used a timeseries analysis to observe that the model-based contribution to starting point and drift rate were negatively correlated within each subject, suggesting that experience within each task, as well as task complexity, mediates the relative contribution of online and offline planning. Taken together, these results suggest that while planning without decision-time deliberation (offline planning) suffices for tasks with low SSC, online planning becomes more necessary with increasing SSC, and that our task and model could be a framework for further investigation of human online planning.

Keywords: Reinforcement learning, Planning, Model-based decision making, Online planning

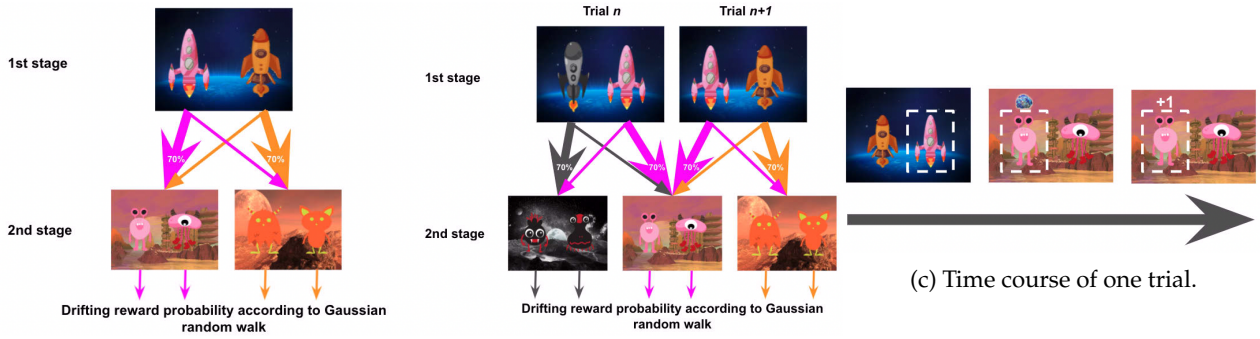
Acknowledgements

This work was supported by NIMH P50MH096889 and NIA R21AG072673 to AMB.

1 Introduction

Planning is one of the most fundamental functions for not only natural but also artificial agents' learning and decision-making. In the realm of reinforcement learning (RL), planning refers to deciding with internal models - a transition function and a reward function - of the environment, thereby being interchangeably referred to as model-based (MB) RL. Utilizing the knowledge of the model to look forward avails efficiency and flexibility of making decisions but comes at a computational cost. Model-free (MF) RL could be seen as a system that is opposite to planning in that it is computationally cheap but inflexible to adapting changes.

MB and MF systems have been found to complement each other in RL. The standard task for measuring this relationship is the two-stage task (TST) that consists of two consecutive decisions, where first-stage decisions stochastically lead to the second-stage decisions [1, 2]. Since the reward is given in the second-stage decision, the best decision to take in the first stage involves using the information of the second stage - in other words, the model. In this framework, a person's tendency to plan can be captured into a parameter w that directly weighs the tendency to use MB vs. MF value, defined by the following formula, $Q = wQ_{MB} + (1-w)Q_{MF}$, where $w = [0, 1]$. Here, MB Q-values (Q_{MB}) are derived by multiplying the transition probability and second-stage values, and MF Q-values (Q_{MF}) are updated via temporal difference (TD) learning upon direct experience. However, a distinction left uninterrogated by this approach is whether plans are constructed *offline* – that is, prior to the trial start – or *online*, after the presentation of options in the moment.



(a) Canonical TST (SSC=2).

(b) Multinomial TST (SSC 3).

(c) Time course of one trial.

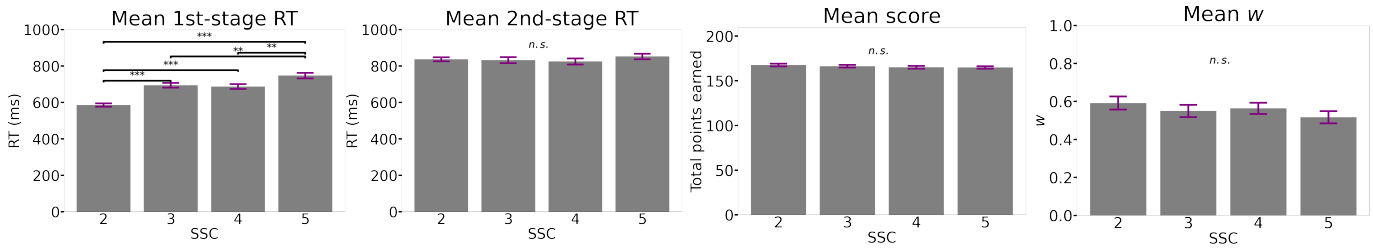
Figure 1: Experimental paradigm. (Figures 1a-1b) Each experiment consists of 300 trials and has 5 catch trials at random sequence. (Figure 1a) Experimental design of the canonical TST (equivalent to SSC-2). Note that the same first-stage state appeared on every trial, which allows for precomputation of plans in the intertrial interval (ITI). (Figure 1b) Experimental design of our proposed multinomial TST. The number of SSC corresponds to the number of rockets in the 1st stage, where each rocket mainly leads to its associated 2nd stage (planet). Therefore, each 1st-stage state varies according to the combination of the rockets. (Figure 1c) Temporal order of TST. Both the first and second stage decisions have a 2-second response window, followed by a feedback phase of 1 second. ITI lasted for 2 seconds with a fixation cross.

This distinction is materially relevant in multiple domains. Recently, it has been shown that artificial MB RL agents rely heavily on offline planning vs. online planning, likely because they can benefit from very large numbers of training examples that can be acquired without direct experience [4]. However, it remains an open question the degree to which humans can similarly eschew online planning, given that they tend to make decisions based on fewer experiences. Eye- and mouse-tracking studies using TST suggest that subjects plan before the trial onset [5, 6], thereby making the direct comparison between online vs. offline planning elusive within the TST paradigm. Here, we propose a novel variant of the TST, the *multinomial* TST, to delineate behavioral patterns of online planning by time-locking the availability of decision-relevant information to the first-stage stimulus onset [3]. This is achieved by increasing the number of possible first-stage options, and selecting combinations of them randomly at each trial, thereby making it difficult for agents to predict and plan the first-stage decisions beforehand. We increase the state-space complexity (SSC) in three levels (3, 4, and 5 first-stage options; Figure 1b), and compare it with the canonical TST which has 2 first-stage options (Figure 1a). Like the original TST, two options are presented on first- and second-stage decisions and the first-stage option stochastically leads to the second-stage state associated with the unchosen first-stage option. We recruited 110 participants between age 18-40 each for every variant - 2-, 3-, 4-, and 5-SSC tasks - via Amazon Mechanical Turk. We excluded 8, 4, 5, and 5 subjects from each experiment, due to one of the following criteria: responding with the same key on more than 95% of the trials, responding implausibly fast (RT below 150 ms) on more than 10% of the trials [9], choosing more than 90% to either option, failing to respond within the response window on more than 20% of the trials, below chance-level model fit, pressing certain button or choosing certain option consecutively for more than 10% of the trials, responding with RT of 0 for more than 5 trials (this indicates that the button was pressed before the onset of the trial), and scored less than 2 out of 5 catch trials. We hypothesized that increasing SSC will evoke online planning; behaviorally, we expected this to

be reflected in an increase in participants' response times (RT) in the first-stage decision, since precomputing the plans before trial onset would be more difficult, and an increased dependency in using trial-specific MB-information on each decision. Our main measures of interest to compare behavioral patterns across SSC come from two models - w from the choice model used by previous TST studies [1, 2] and the non-decision time (NDT; t), drift rate (v), and starting-point bias (z) parameters in a reinforcement learning diffusion decision model (RLDDM; [8]), here extended to incorporate the multiple value timeseries (MB and MF) available in the TST. Specifically, we measure online planning with the MB evidence accumulation rate (v_{MB}), since the drift rate stands for evidence accumulated at each given timepoint. In contrast, we hypothesize that the MB evidence already accumulated before the start of the trial (z_{MB}) represents offline planning.

2 Results

First, we tested the hypothesis that complex environments recruit online decision evaluation by analyzing the response time of first-stage states as a function of SSC (Figure 2). The first-stage RT indeed increased as a function of SSC; a one-way ANOVA yielded a significant difference between SSC ($F_{(3,412)} = 26, p < .001$), and a post-hoc t-test revealed that the RT of SSC-2 are significantly lower than other conditions (SSC-2 vs. SSC-3: $t_{(203)} = -6.69, p < .001$, SSC-2 vs. SSC-4: $t_{(201)} = -6.17, p < .001$, SSC-2 vs. SSC-5: $t_{(205)} = -9.01, p < .001$), and 1st-stage RT in SSC-5 were also significantly slower than other conditions (SSC-3 vs. SSC-5: $t_{(210)} = -2.66, p = .009$, SSC-4 vs. SSC-5: $t_{(208)} = -2.98, p = .003$; Figure 2a). It is also notable that neither the mean score (Figure 2c) nor mean w (Figure 2d) are different across conditions (all $p > 0.3$). This suggests that the differences in first-stage RT cannot be explained by the difficulty of the task *per se* or the degree of model use in each task. Further evidence in favor of this point is that second-stage RT, which also involves binary choice but not planning into the future, was invariant among conditions ($F_{(3,412)} = 1.86, p > .13$; Figure 2b).



(a) Mean first-stage RT (b) Mean second-stage RT (c) Mean score (d) Mean w

Figure 2: Behavioral analyses across conditions. Error bars indicate standard error. Significant results of pairwise t-tests are indicated with asterisks: $* = p < .05, ** = p < .01, *** = p < .001$) (a-b) Mean first- and second-stage RT across different SSC. Prior to statistical tests such as one-way ANOVA and t-tests, each participant's mean RTs for 1st- and 2nd-stage decisions were log-transformed and normalized. (a) First-stage RT of SSC-2 was significantly shorter than the rest conditions, while the opposite was true for SSC-5. There was no significant difference between 1st-stage RT of SSC-3 and SSC-4 ($t_{(206)} = .38, p > .7$). (b-d) A one-way ANOVA indicated no significant difference between groups in mean second-stage RT, mean score, and mean w across groups (all $p > .13$).

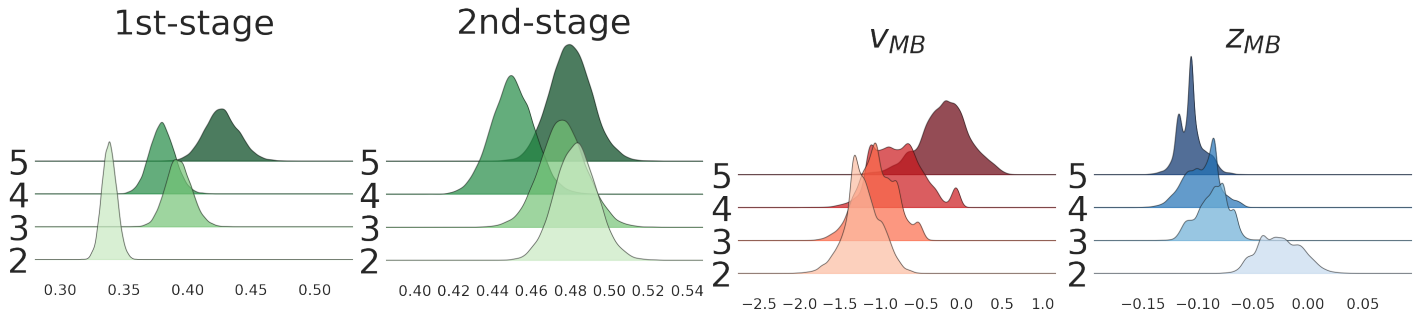
Next, we analyzed the data using an RLDDM to identify the within-trial pattern of planning. Our model inherited features of the original RLDDM [8], but augmented to model the TST. Specifically, in order to delineate the use of the MB vs. MF values at each 1st-stage decision, we added additional parameters to the original drift rate and the starting-point bias reflecting the trial-by-trial variation in these quantities, according to the following specification:

$$v = v_0 + v_{MB}\delta(Q_{MB}) + v_{MF}\delta(Q_{MF}) + v_{int}\delta(Q_{MB})\delta(Q_{MF}) \quad (1)$$

$$z = z_0 + z_{MB}\delta(Q_{MB}) + z_{MF}\delta(Q_{MF}) + z_{int}\delta(Q_{MB})\delta(Q_{MF}) \quad (2)$$

, where $\delta(Q_{MB})$ and $\delta(Q_{MF})$ stand for the MB or MF Q-value differences of the two first-stage options, respectively. In our RLDDM, the decision threshold (a) was fixed to 1 for a straightforward interpretation of v and z, and a sigmoid function was applied to z so that the starting point was bounded to $[0, 1]$. The model was fitted to the data via the HDDM package [10], where each SSC was estimated by 5 chains of 15,000 samples with 2,000 burn-in samples.

In line with the overall RT, the 1st-stage NDT significantly increased as a function of SSC ($F_{(3,412)} = 14.03, p < .001$; mean $t_{SSC-2} = .34$ s, mean $t_{SSC-3} = .4$ s, mean $t_{SSC-4} = .39$ s, mean $t_{SSC-5} = .43$ s; Figure 3a), while the 2nd-stage NDT did not differ among conditions (Figure 3b). This may reflect that in order to plan online in higher SSC environments, participants must spend a longer time identifying the options present on the screen (ostensibly to establish the current



(a) Posterior distribution of first-stage NDT (t)

(b) Posterior distribution of second-stage NDT (t)

(c) Posterior distribution of v_{MB}

(d) Posterior distribution of z_{MB}

Figure 3: Posterior distribution of parameters estimated from RLDDM. For each SSC, the RLDDM model was estimated via HDDM on 5 chains that iterated for 15,000 samples, of which 2,000 were burn-in samples. The results show distribution of samples concatenated from all 5 chains. (a-b) The unit for NDT (t) is seconds.

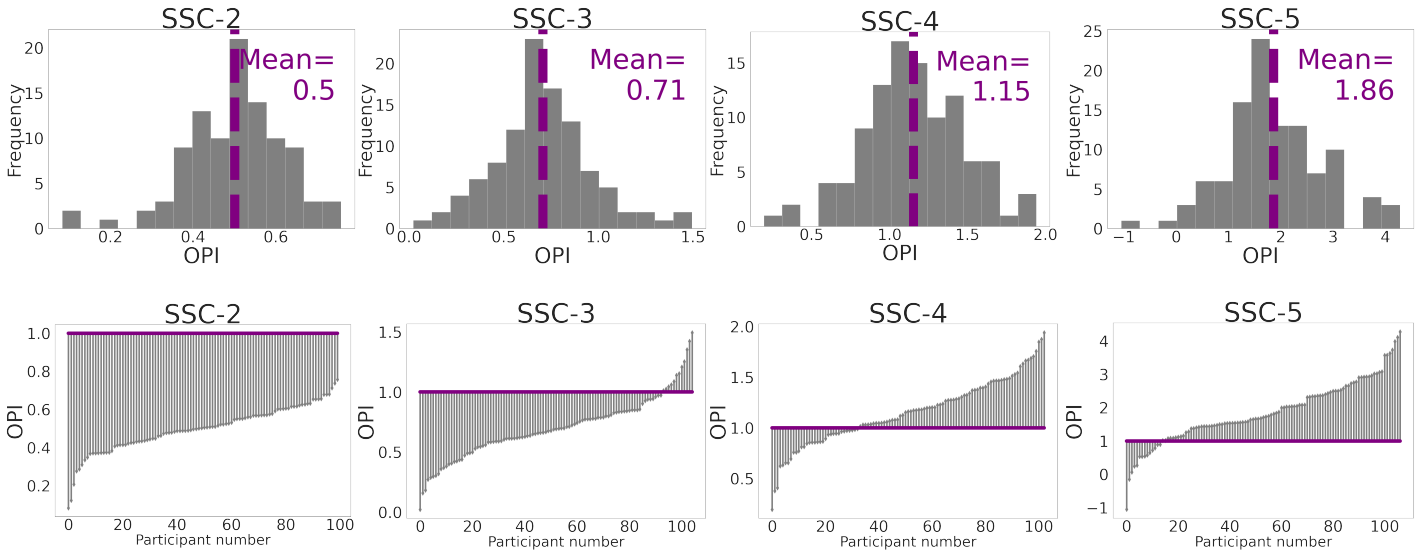


Figure 4: The distribution of the online planning index (OPI), defined by the ratio of v_{MB} to z_{MB}. (First row) The histogram of OPI for each condition. (Second row) A stem plot of OPI for each condition, with OPI=1 as the baseline.

decision tree) at each trial, which would have been unnecessary for the original TST. Importantly, after accounting for this condition-wise shift in RT using the NDT parameter, we examined how trial-varying MB and MF values affected both the starting point (z) and drift rate (v) differentially in each condition. Supporting the hypothesis that participants' dependency on using online, relative to offline, planning should increase as a function of SSC, we found a significant increase of v_{MB} and decrease of z_{MB} as a function of SSC through one-way ANOVA ($F_{(3,412)} = 2373.11, p < .001$; Figure 3c; $F_{(3,412)} = 55.18, p < .001$; Figure 3d). Post-hoc pairwise t-tests revealed that while all distributions of v_{MB} were significantly different among SSC (all $p < .001$), all pairs but SSC-4 vs. SSC-5 were significantly different for z_{MB} (rest $p < .001$).

Next, we examined whether online and offline planning traded off within each subject. Specifically, we calculated an *online planning index* (OPI), given by $OPI = (1 - z_{MB})/(1 - v_{MB})$. The first and second row of Figure 4 show that as SSC increases, so does the proportion of subjects adopting online planning ($F_{(3,412)} = 140.61, p < .001$), where subjects in SSC-2 and SSC-3 heavily rely on offline planning while the majority of participants in SSC-4 and SSC-5 uses online planning ($OPI = 1$ indicates equal use of online and online planning). To determine the extent to which general planning could be explained by online planning, we correlated w from the choice model and OPI from the RLDDM. Again supporting the hypothesis that online planning becomes more relevant as SSC increases, only SSC-5 showed a significant correlation between the two parameters (Pearson's $r = .25, p = .009$), indicating that planning in SSC-5 is heavily driven by online planning. Finally, to examine the tradeoff between online and offline planning over the course of task experience, we performed a sliding-window RLDDM with a window size of 50 trials to obtain the timeseries of v_{MB} and z_{MB}. We then correlated the timeseries of v_{MB} and z_{MB} for each subject, and (Figure 5) shows that all SSC yields significant, medium to strong negative correlations between v_{MB} and z_{MB}. Entering the z-transformed correlation coefficients into pairwise t-tests against zero yielded significant differences for all conditions (all $p < .001$).

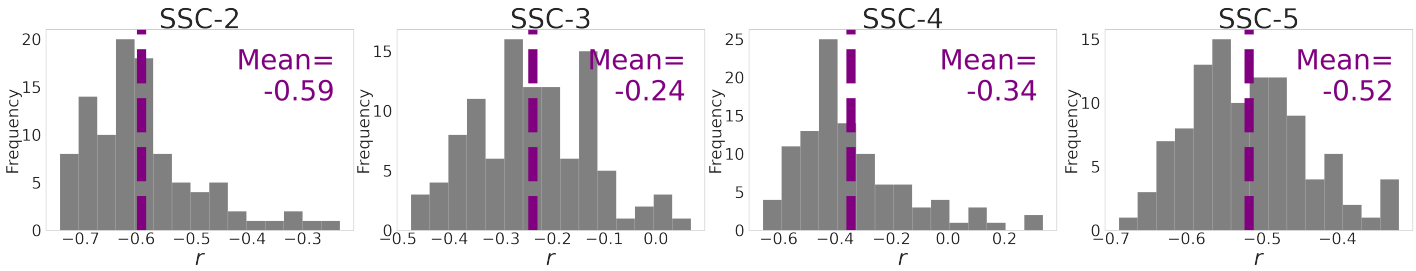


Figure 5: The distribution of the correlation coefficient (Pearson's r) between the timeseries of v_{MB} and z_{MB} for each subject. The timeseries is generated by applying the RLDDM model to a window of 50 trials, and sliding the start of the window from the first trial to the 250th trial.

3 Conclusion

The TST framework has been an important method for investigating planning, but a key limitation of this framework is that it prevents evaluating the online deliberation process directly in behavior. Here, we introduce a multinomial TST which sets aside this limitation, and across four experiments, demonstrate that response time patterns reflect a sensitivity to both each task's differing state complexity and also trial-wise learned values. The preliminary results reported here suggest that our novel multinomial TST can provide a useful foundation for investigating online planning patterns in humans. Importantly, we have shown that although plans could be made in advance in simplistic environments, this is gradually hampered with complexity and thus recruits decision-time planning. This trade-off relationship between online and offline planning is observed across both SSC and experience within the task. Together with this trade-off, the correlation between OPI and w in the most complex environment may provide an explanation for the invariant general planning (w) across contexts, also shown in previous studies [7]. Also, OPI, which directly compares the relationship between the two planning mechanisms, could be used as a new individual-difference measure with potential clinical relevance. We also anticipate that bringing this decision-time activity under behavioral control and experimental manipulation will allow us to interrogate the algorithmic structure of model-based decisions, to understand the balance between online and offline control in humans performing novel tasks, and will allow for evaluating online deliberation activity using neuroimaging measures.

References

[1] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[2] Johannes H Decker, A Ross Otto, Nathaniel D Daw, and Catherine A Hartley. From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, 27(6):848–858, 2016.

[3] Laura Fontanesi, Sebastian Gluth, Mikhail S Spektor, and Jörg Rieskamp. A reinforcement learning diffusion decision model for value-based decisions. *Psychonomic bulletin & review*, 26(4):1099–1121, 2019.

[4] Jessica B Hamrick, Abram L Friesen, Feryal Behbahani, Arthur Guez, Fabio Viola, Sims Witherspoon, Thomas Anthony, Lars Buesing, Petar Veličković, and Théophane Weber. On the role of planning in model-based deep reinforcement learning. *arXiv preprint arXiv:2011.04021*, 2020.

[5] Arkady Konovalov and Ian Krajbich. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. *Nature communications*, 7(1):1–11, 2016.

[6] Arkady Konovalov and Ian Krajbich. Mouse tracking reveals structure knowledge in the absence of model-based choice. *Nature communications*, 11(1):1–9, 2020.

[7] Wouter Kool, Samuel J Gershman, and Fiery A Cushman. Planning complexity registers as a cost in metacontrol. *Journal of cognitive neuroscience*, 30(10):1391–1404, 2018.

[8] Mads L Pedersen and Michael J Frank. Simultaneous hierarchical bayesian parameter estimation for reinforcement learning and drift diffusion models: a tutorial and links to neural data. *Computational Brain & Behavior*, 3(4):458–471, 2020.

[9] Nitzan Shahar, Tobias U Hauser, Michael Moutoussis, Rani Moran, Mehdi Keramati, Nspn Consortium, and Raymond J Dolan. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology*, 15(2):e1006803, 2019.

[10] Thomas V Wiecki, Imri Sofer, and Michael J Frank. Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, page 14, 2013.