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

Word count: 4,993

This research was supported by the Australian Government through the Australian Research Council's Discovery Project funding scheme (project DP180103613, to GEH and SDB) and Discovery Early Career Award funding scheme (project DE170100177, to GEH). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

The raw data, experimental code, and analysis code for this article can be found online at `https://osf.io/qg326/` .

*To whom correspondence may be addressed (nathan.j.evans@uon.edu.au).

Abstract

Theories of perceptual decision-making have been dominated by the idea that evidence accumulates in favor of different alternatives until some fixed threshold amount is reached, which triggers a decision. Recent theories have suggested that these thresholds may not be fixed during each decision, but change as time passes. These collapsing thresholds can improve performance in particular decision environments, but reviews of data from typical decision-making paradigms have failed to support collapsing thresholds. We designed three experiments to test collapsing threshold assumptions in decision environments specifically tailored to make them optimal. An emphasis on decision speed encouraged the adoption of collapsing thresholds – most strongly through the use of response deadlines, but also through instruction to a lesser extent – but setting an explicit goal of reward rate optimality through both instructions and task design did not.

Keywords:

Decision-making — Collapsing thresholds — Diffusion model — Bayesian hierarchical modelling

Introduction

Modern theories of rapid decision-making are dominated by a class of formalized process models known as evidence accumulation models (EAMs; Stone, 1960; for modern reviews see Donkin & Brown, 2018, or Ratcliff, Smith, Brown, & McKoon, 2016). EAMs have been foundational to our understanding of decision-making across a wide variety of contexts, ranging from simple perceptual tasks to complex discrete choice data (Hawkins et al., 2014), stop-signal paradigms (Matzke, Dolan, Logan, Brown, & Wagenmakers, 2013), go/no-go tasks (Gomez, Ratcliff, & Perea, 2007), absolute identification (Brown, Marley, Donkin, & Heathcote, 2008), optimality studies (Starns & Ratcliff, 2012; Evans & Brown, 2017), personality traits (Evans, Rae, Bushmakin, Rubin, & Brown, 2017), neural data (Forstmann et al., 2011), and clinical populations (Ho et al., 2014). EAMs propose that decisions involve steadily accumulating evidence in favor of the various choice alternatives until a sufficient quantity of evidence has been accumulated for one of the alternatives to reach a pre-determined decision threshold, which triggers a response. Decision thresholds are assumed to be under the strategic control of the decision-maker. For example, to make slower and more careful decisions one can set a higher threshold on the accumulated evidence.

The most widely used EAM is the diffusion model (Stone, 1960; Ratcliff, 1978; Ratcliff & Rouder, 1998), which proposes that the evidence accumulation process occurs between two directly opposing alternatives, illustrated in Figure 1. The rate of evidence accumulation is called the "drift rate", and the diffusion model also allows for potential response biases (the "starting point") and time spent in non-decision related processes ("non-decision time"). This basic version of the diffusion model has been extended to account for key qualitative benchmarks of response time data (e.g., the relative speed of correct vs. incorrect decisions; Ratcliff, 1978; Ratcliff & Rouder, 1998), by the inclusion of

between-decision variability in some model parameters.

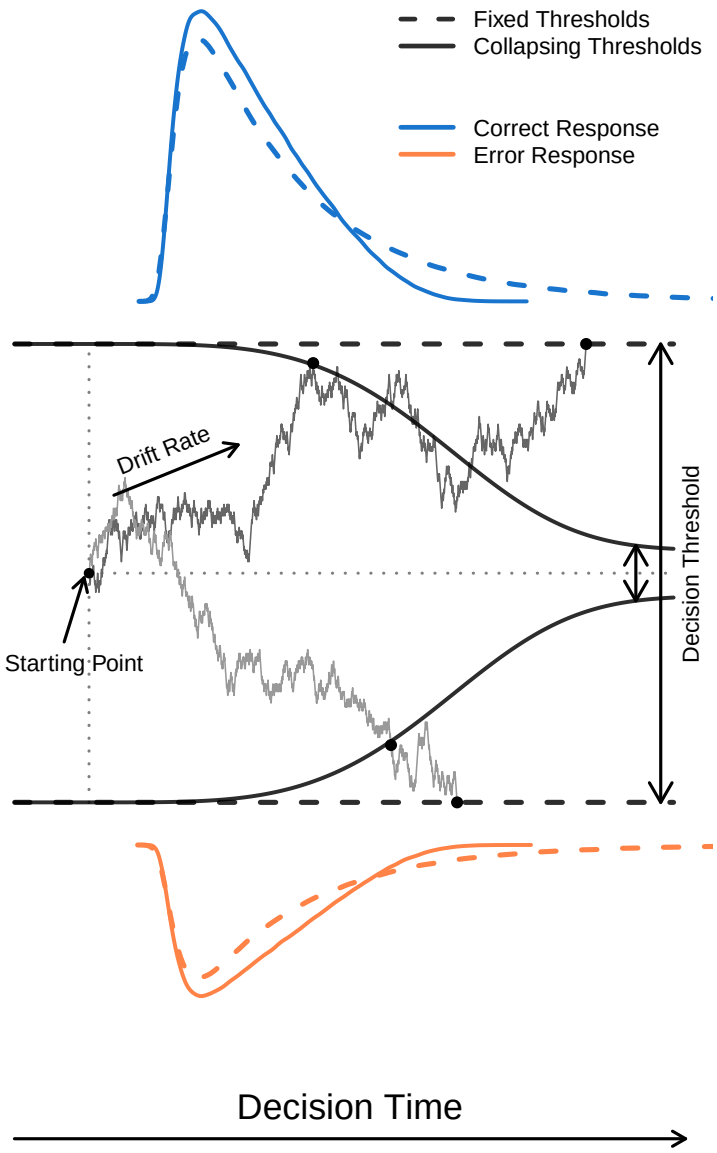

Figure 1. The diffusion model framework, with fixed or collapsing thresholds (dashed and solid lines, respectively). For particular values of the drift rate, starting point and non-decision time, when the two models reflect equally cautious decisions at stimulus onset – equal height decision thresholds for the very fastest decisions, as in this example – the collapsing thresholds model predicts equal or faster response times than the fixed thresholds model. This can be seen in the predicted response time distributions for correct (upper, blue) and incorrect (lower, orange) responses of the two models. The key trend in data that discriminates the two models is that the long, positively skewed tail predicted by the fixed thresholds model is partially truncated by the collapsing thresholds model.

Despite the successful history of the diffusion model in accounting for data, a number of recent studies have proposed that a fundamental assumption of the model may be incorrect: the decision threshold may not be constant over the course of a decision. Instead, these studies have suggested that decision thresholds might decrease with passing time, a proposition commonly referred to as "collapsing thresholds" (Cisek, Puskas, & El-Murr, 2009; Ditterich, 2006a; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012; Thura, Beauregard-Racine, Fradet, & Cisek, 2012). Collapsing thresholds imply that as a decision takes longer and longer, the evidence required to trigger a response reduces (compare the solid black decision thresholds with the dashed black decision thresholds in Figure 1).

There have been two primary theoretical motivations provided in favor of collapsing thresholds. Firstly, collapsing thresholds can lead to improved performance in terms of "reward rate". Reward rate is the expected number of correct decisions that can be made in some time period, or equivalently, the expected time between correct decisions. Maximizing reward rate requires carefully balancing caution and urgency. Too-urgent decision-making will lead to fewer correct decisions, while too-cautious decision-making will lead to fewer decisions made in total. Wald and Wolfowitz (1948) showed that a diffusion model with fixed thresholds optimizes reward rate when all the decisions encountered are of equal difficulty. If decisions vary unpredictably in difficulty from one to the next – a common design in the study of rapid decision-making – then adopting collapsing thresholds can lead to a higher reward rate than is possible using even the best possible fixed thresholds (Drugowitsch et al., 2012; Thura et al., 2012). This is because collapsing thresholds allow the decision-maker to capitalize on the advantages of high *and* low fixed thresholds: to maximize accuracy in easy decisions with high drift rates as thresholds are initially high and few errors are made due to randomness in the process, and expending little time on

harder trials with low drift rates as the thresholds become lower as the decision continues to unfold.

The second primary motivation for collapsing thresholds is that they allow the decision maker to maintain good performance when fast decisions matter more than accurate decisions. Experimentally, this usually takes the form of explicit instructions that emphasize the speed of responding or, less commonly, deadline manipulations that restrict the time available to make decisions. Speeded conditions – whether via instructions or deadlines – require participants to avoid committing too much time to a single decision, which leaves the decision-maker with two potential strategies: a low fixed threshold, or a collapsing threshold. A low fixed threshold is quickly reached, even in difficult decisions when the drift rate is low, which satisfies the speed goal. However, the low threshold also leads to many incorrect decisions, and does not provide a principled way of avoiding misses. Collapsing thresholds can improve accuracy in easier trials by setting initially high thresholds, allowing participants to achieve an acceptable level of overall accuracy (even if it does not lead to the maximization of reward rate). Collapsing thresholds also maintain quick performance on trials that take longer periods of time, and provide a clear, theoretically satisfying mechanism to ensure that a decision is made before time expires in deadline tasks, by dynamically decreasing the amount of evidence required to trigger a decision to zero prior to the occurrence of the deadline (Frazier & Yu, 2007).

Although collapsing thresholds have been theoretically motivated on grounds of reward rate maximization and enhancing speeded performance, comparisons of fixed and collapsing threshold models to date have mostly failed to examine task performance in these decision contexts. Initial comparisons focused on non-standard experimental tests that aimed to qualitatively discriminate between the theories, and these found evidence in favor of collapsing thresholds or the related "urgency signal" (Cisek et al., 2009; Dit-

terich, 2006a; Drugowitsch et al., 2012; Thura et al., 2012). However, the value of such qualitative comparisons between theories can be limited by a focus on small subsets of the data (e.g., just one special condition of a wider experiment; Cisek et al., 2009; Thura et al., 2012) or on broad summary statistics (e.g., mean response time and accuracy; Thura et al., 2012; Winkel, Keuken, van Maanen, Wagenmakers, & Forstmann, 2014). Qualitative comparisons often also require the theories to be replaced by simplified or incomplete versions, to aid researchers' intuitions, and this can limit the relevance of the comparisons to the actual theories being tested (see Evans, Hawkins, Boehm, Wagenmakers, & Brown, 2017). More recent studies have compared complete versions of the models against full data sets, using data from typical decision-making paradigms. Those investigations have found fixed thresholds to provide a better explanation of human decision-making (Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Hawkins, Wagenmakers, Ratcliff, & Brown, 2015; Voskuilen, Ratcliff, & Smith, 2016), while collapsing thresholds provide a better account of non-human primate data (Hawkins, Forstmann, et al., 2015; though see Evans & Hawkins, 2019, for evidence that this discrepancy might be due to differences in experimental procedure across species).

An important remaining question is whether people adopt collapsing thresholds when they find themselves in those decision contexts where collapsing thresholds are theoretically superior. Situations that require reward rate maximization or emphasize fast performance can result in advantages for collapsing thresholds over fixed thresholds. However, previous quantitative comparisons between the models have not focused on tasks in which participants were instructed to achieve either of these goals. Most often, participants have not been instructed to maximize their reward rate, and have also been given neutral (or absent) instructions about emphasizing caution or urgency. When given such neutral instructions, people tend to emphasize the accuracy rather than the speed of their decisions (Forstmann

et al., 2008), leading to performance that is more cautious than optimal (Evans & Brown, 2017; Starns & Ratcliff, 2012; Evans, Bennett, & Brown, 2018). This could have biased previous investigations in favor of fixed thresholds over collapsing thresholds, as the latter are most useful when speed is emphasized.

With data from three experiments, we assessed whether people adopted collapsing thresholds in the situations where those thresholds are most adaptive. In Experiment 1, we used decisions which varied unpredictably in difficulty, explicitly instructed the participants to maximize reward rate, and provided them with detailed feedback (Evans & Brown, 2017; Evans, Bennett, & Brown, 2018). Experiments 2 and 3 investigated two different ways to emphasize decision speed; a decision deadline in Experiment 2 and speed-emphasis instructions in Experiment 3. Across all experiments, we carried out model estimation using a Bayesian hierarchical approach, which takes into account the uncertainty in parameter estimates and allows inferences to be drawn simultaneously at the group and individual-participant levels. Our primary analyses compare the fixed and collapsing threshold models for each participant using the Deviance Information Criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), which accounts for functional form complexity (see Evans & Brown, 2018, Myung, 2000, and Evans, Howard, Heathcote, & Brown, 2017 for more details on functional form complexity). All models were estimated using analytic ("closed form") probability density functions. This decreases a potential source of variability when compared to analyses based on model simulation (Hawkins, Forstmann, et al., 2015; Evans, Hawkins, et al., 2017), because simulated-based methods can result in error in the estimated posterior distributions and model selection metrics (Holmes, 2015).

Experiment 1

Method

Participants. Sixty-three undergraduate students from the University of Newcastle completed the experiment online, which was approved by the University of Newcastle Human Research Ethics Committee, and were reimbursed with course credit. Participants completed the experiment on a web browser interface at a time and location of their choosing, with the experiment delivered through purpose-built Javascript code. Prior to commencing data collection, we defined an exclusion criterion for the task based on decision accuracy of 60%, where participants scoring below this criterion would not be considered to have performed the task correctly (though it should be noted that we did not formally preregister this exclusion criterion). This exclusion criterion was based on the similar experimental paradigms of Evans and Brown (2017) and Evans, Bennett, and Brown (2018), but made slightly more lenient (i.e., 60% instead of the 70% of Evans & Brown, 2017 and Evans, Bennett, & Brown, 2018) due to the greater overall difficulty (i.e., lower overall dot movement coherence) of this experiment. Once the data had been collected, all participants who fell below this criterion were removed, resulting in removal of data from six participants. In supplementary material, we also demonstrate that analyses which include these under-performing participants lead to an identical overall pattern of results. The sample size for this experiment and the two subsequent experiments were based upon previous similar investigations.

Task and procedure. Participants made decisions about apparent motion in random dot kinematograms (Roitman & Shadlen, 2002; Evans & Brown, 2017), which is a standard task in decision-making studies. Our stimuli used the white-noise algorithm (Pilly & Seitz, 2009) with 40 white dots on a black background. On each frame (66.7ms) each of the dots

moved. Some of the dots moved "coherently" – in the same direction as each other – and the others moved random distances in random directions. The direction of the coherent movement was randomly selected from either top-left or top-right motion for each trial, in equal proportions. Participants were tasked with determining whether the dots moved toward the top-left or the top-right of the screen, by pressing either the "z" or "/" keys, respectively. Each dot was 3 pixels in diameter, and all dots always remained within a radius of 75 pixels at the center of the display; any dot that moved outside of the central radius was replaced within the circle on the subsequent frame, at a random location. On each frame, the size of the movement for each coherently moving dot was $\sqrt{18}$ pixels: 3 pixels up, and 3 pixels towards either the left or the right of the screen, depending on the dot direction selected for that trial. Participants received feedback after each trial: correct feedback was displayed for 300ms, incorrect feedback was displayed for 800ms, and responses that were faster than 250ms were discouraged by following them with a 1,500ms timeout, as such responses are faster than basic perceptual processes and thus cannot reflect a decision about the stimulus information.

To create the emphasis on reward rate, we used the two methods used previously within the reward rate optimality literature (Evans & Brown, 2017; Evans, Bennett, & Brown, 2018; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Simen et al., 2009; Starns & Ratcliff, 2012, 2010; Balci et al., 2011). Firstly, participants were instructed at the beginning of the task that their goal was to make as many correct responses (i.e., gain as many rewards) as possible. Secondly, there was a fixed amount of time available to make decisions in each block, which means that maximizing reward rate and maximizing the number of correct responses were equivalent decision strategies. In general, this structure differs to the majority of cognitive psychology experiments that have a fixed number of trials in each block, in which case the two decision strategies are not equivalent (e.g.,

see Hawkins, Brown, Steyvers, & Wagenmakers, 2012). Participants were explicitly made aware of this task structure, and were informed that it implied that the speed at which they performed the task would affect how many trials they experienced, and therefore, the number of rewards they could possibly achieve. There were 30 blocks, each 1 minute in duration. Additionally, at the end of each block except the first three, participants were given feedback on their reward rate over the last 200 trials, or all trials completed to date if they had not yet completed 200 trials. This kind of feedback assists people in maximizing reward rates (Evans & Brown, 2017; Evans, Bennett, & Brown, 2018).

Design and data analysis. Experiment 1 manipulated a single within-subjects factor: motion coherence of the random dot kinematogram, manipulated across 4 levels – 0%, 5%, 10%, and 40% coherence (equivalently, 0, 2, 4, or 16 coherently moving dots). This manipulates the difficulty of decisions, because it is harder to identify the motion direction when coherence is lower.

Data from the first 21 blocks of trials were removed from analysis. This was to allow participants time to learn from the feedback about reward rate performance, to improve their performance and adopt optimal (or close-to-optimal) strategies where possible. Previous work with the same block-to-block feedback as this experiment showed that participants required around 15 blocks to settle on a threshold (Evans & Brown, 2017; Evans, Bennett, & Brown, 2018). In the supplementary materials we demonstrate that this exclusion led to performance which was close to stationary over the included blocks. Block-to-block feedback on reward rate was not used in Experiments 2 and 3, so only the very first block was excluded in those experiments. We also excluded responses faster than 250ms or slower than 5,000ms. The average number of trials/block across participants in the included blocks (22-30) was 42.6, resulting in approximately 380 trials per participant for analysis.

Our starting point for developing a quantitative implementation of the fixed and collapsing thresholds models was the basic (also referred to as "simple") diffusion model of Stone (1960), defined by four parameters: drift rate, decision threshold, starting point of evidence accumulation, and the time consumed by non-decision related processes. Our fixed threshold diffusion model was the "full" diffusion model, which included three additional parameters reflecting trial-to-trial variability in drift rate, starting point, and non-decision time. We also tested the "simple" version of the fixed thresholds diffusion model, which does not assume trial-to-trial variability in parameters, though in most cases this provided a poorer DIC value than the "full" diffusion model. In the few cases where it did not, there were no changes to the overall qualitative pattern of results. For brevity, we do not further discuss the "simple" diffusion model in the main text, and refer the reader to supplementary materials for complete details of the simple diffusion model analysis.[1]

We allowed drift rate to vary across the difficulty manipulation of the number of coherent dots in the trial, meaning that each participant had 10 free parameters for the fixed threshold diffusion model (with the stochastic within-trial variability in drift rate fixed to 1 to satisfy a scaling property of the model). We assumed a hierarchical structure so that the free parameters for each individual were constrained by a group-level distribution for that parameter. Formally, the fixed thresholds diffusion was defined as:

[1]The analyses reported in the supplementary materials are based on methods developed by several groups in last five years: (Evans & Brown, 2017; Evans, 2019b; Evans, Bennett, & Brown, 2018; Evans, Brown, Mewhort, & Heathcote, 2018; Hawkins, Forstmann, et al., 2015; JASP Team, 2018; Lerche, Voss, & Nagler, 2017; Palestro, Weichart, Sederberg, & Turner, 2018; Rae, Heathcote, Donkin, Averell, & Brown, 2014; Rouder, Morey, Speckman, & Province, 2012).

$$
\begin{aligned}
&\textit{Data level}: \\
(RT_i, resp_i) &\sim Diffusion(v_{coherence,i}, z_i, ter_i, \\
&\qquad a_i, s_{v,i}, s_{z,i}, s_{ter,i})
\end{aligned}
$$

$$
\begin{aligned}
&\textit{Group level}: \\
v_{coherence,i} &\sim N(\mu_{v,coherence}, \sigma_{v,coherence}) \\
\frac{z_i}{a_i} &\sim TN(\mu_z, \sigma_z, 0, 1) \\
ter_i &\sim TN(\mu_{ter}, \sigma_{ter}, 0, Inf) \\
a_i &\sim TN(\mu_a, \sigma_a, 0, Inf) \\
s_{v,i} &\sim TN(\mu_{s_v}, \sigma_{s_v}, 0, Inf) \\
s_{z,i} &\sim TN(\mu_{s_z}, \sigma_{s_z}, 0, Inf) \\
s_{ter,i} &\sim TN(\mu_{s_{ter}}, \sigma_{s_{ter}}, 0, Inf)
\end{aligned}
$$

$$
\begin{aligned}
&\textit{Prior distributions}: \\
\mu_{v,coherence} &\sim N(3, 3) \\
\mu_z &\sim TN(.5, .5, 0, 1) \\
\mu_{ter} &\sim TN(.3, 1, 0, Inf) \\
\mu_a &\sim TN(2, 2, 0, Inf) \\
\mu_{s_v}, \mu_{s_z}, \mu_{s_{ter}} &\sim TN(1, 1, 0, Inf) \\
\sigma_{v,coherence}, \sigma_a &\sim \Gamma(1, 1) \\
\sigma_z, \sigma_{ter}, \sigma_{s_v}, \sigma_{s_z}, \sigma_{s_{ter}} &\sim \Gamma(.5, .5)
\end{aligned}
$$

where i denotes participants, $\sim$ means "is distributed as", TN indicates a truncated normal distribution with parameters (in order): mean, standard deviation, minimum, and maximum. Γ refers to the gamma distribution with parameters shape and scale, and μ and σ refer to the mean and standard deviation of the group-level distribution, respectively. Regarding the specific parameters of the model, a refers to the distance between the decision thresholds, v refers to the drift rate and s_v refers to the between-trial variability in drift rate, z refers to the starting point and s_z refers to the between-trial variability in starting point, and ter refers to the non-decision time and s_{ter} refers to the between-trial variability in non-decision time.

The collapsing threshold model started with the "simple" diffusion model and added a collapsing Weibull function for the threshold. This added two free parameters, for the shape and scale of the Weibull function (for details, see Hawkins, Forstmann, et al., 2015). From their initial values, the thresholds collapsed together to meet at a point equal to the starting point of the evidence accumulation process. For numerical stability, we allowed a numerically insignificant difference between the end points of the thresholds' collapse, 10^{-3}. As for the fixed threshold model, drift rate varied with the number of coherent dots. These assumptions imply that each participant had 9 free parameters under the collapsing threshold diffusion model (with the stochastic within-trial variability in drift rate was fixed to 0.1 to satisfy a scaling property of the model). With an analogous Bayesian hierarchical structure as the fixed thresholds model, the collapsing thresholds diffusion model was formally defined as:

$$
\begin{aligned}
&Data\ level: \\
(RT_i, resp_i) &\sim Diffusion(v_{coherence,i}, z_i, ter_i, \\
&\qquad a_i, shape_i, scale_i)
\end{aligned}
$$

$$
\begin{aligned}
&Group\ level: \\
v_{coherence,i} &\sim N(\mu_{v,coherence}, \sigma_{v,coherence}) \\
\frac{z_i}{a_i} &\sim TN(\mu_z, \sigma_z, 0, 1) \\
ter_i &\sim TN(\mu_{ter}, \sigma_{ter}, 0, Inf) \\
a_i &\sim TN(\mu_a, \sigma_a, 0, Inf) \\
shape_i &\sim TN(\mu_{shape}, \sigma_{shape}, 0, Inf) \\
scale_i &\sim TN(\mu_{scale}, \sigma_{scale}, 0, Inf)
\end{aligned}
$$

$$
\begin{aligned}
&Prior\ distributions: \\
\mu_{v,coherence} &\sim N(.5, 2) \\
\mu_z &\sim TN(.5, .5, 0, 1) \\
\mu_{ter} &\sim TN(.3, 1, 0, Inf) \\
\mu_a &\sim TN(.2, .6, 0, Inf) \\
\mu_{shape}, \mu_{scale} &\sim TN(3, 3, 0, Inf) \\
\sigma_z, \sigma_{ter} &\sim \Gamma(.5, .5) \\
\sigma_a, \sigma_{shape}, \sigma_{scale} &\sim \Gamma(1, 1) \\
\sigma_{v,coherence} &\sim \Gamma(.5, 1)
\end{aligned}
$$

where the notation is as described for the fixed thresholds model with the exception that *shape* refers to the shape parameter of the Weibull collapsing function, and *scale* refers to the scale parameter of the Weibull collapsing function. For mostly historical reasons, the scaling constant in the collapsing thresholds diffusion model was 0.1, compared to a scaling constant of 1 in the fixed thresholds diffusion model. This means that the measurement units of many of the parameters differ between models by a factor of 10. The different prior specifications for the two models cover roughly equivalent a-priori plausible values for all parameters. The priors we have used, for both models, are relatively uninformative, and methods of model selection based on predictive accuracy – such as DIC – only include the model likelihood function (i.e., $p(y|\boldsymbol{\theta})$) in the calculation, meaning that broad priors have little influence on their calculation.

We did not include between-trial variability parameters within the collapsing thresholds model for two key reasons. Firstly, the inclusion of these additional parameters would make the models analytically intractable, meaning that the models would have to be estimated through simulation methods rather than using analytic probability density functions. This adds an extra layer of potential noise in the results. Secondly, previous studies have claimed that collapsing thresholds can qualitatively account for many of the phenomena that between-trial variability parameters were placed within the fixed thresholds diffusion model to capture (Ditterich, 2006b; Palmer, Huk, & Shadlen, 2005; Shadlen & Kiani, 2013). Both the "full" diffusion model and the "simple" collapsing bounds model are already quite complex models which capture a great deal of the fine structure of the data. The debates in the literature about their adequacy are about whether simpler versions might suffice, which is what we test here. Testing an even more complex version, by including between-trial variability parameters in addition to the regular collapsing parameters may result in an overly-flexible model, as both types of parameters are capturing the same aspects of

the data. As noted above, the simple version of the fixed thresholds model did not lead to qualitative changes in the pattern of results, suggesting that omitting (including) the trial-to-trial variability parameters in the collapsing (fixed) thresholds model did not drive our pattern of results.

The likelihood of the data for each set of parameters was obtained through code extracted from the $fastdm$ package (Voss & Voss, 2007) for the fixed thresholds diffusion model, and custom code developed from the solutions of Smith (2000) for the collapsing thresholds diffusion model. Both models allowed for the possibility of responses unrelated to the experiment, by including a contamination process (Ratcliff & Tuerlinckx, 2002). This was a mixture model in which responses were assumed to come from the model, with proportion $(1 - x)$, and to come from a contaminant process, with proportion x. The contaminant process assumed completely independent responses: evenly split over the two response choices, and uniformly distributed over the entire range of observable RTs – up to the exclusion limit. We also tested variants of the models without a contamination process, though in most cases these models provided poorer DIC values than their counterparts with the contamination process, and the inclusion of these non-contamination models did not change the qualitative patterns in the overall selections. Details of these analyses are reported in the supplementary materials. To estimate the posterior distributions, we used Differential Evolution Markov chain Monte Carlo (DE-MCMC: Turner, Sederberg, Brown, & Steyvers, 2013). We ran $3k$ chains, where k is the number of free parameters for each individual participant, which was the greatest number of free parameters estimated in a sampling block. We ran 2,000 iterations for burn-in, which included a migration algorithm implemented every 10 iterations between the 500^{th} and $1{,}500^{th}$ iteration, with convergence assessed through visual inspection, and then drew 1,500 samples from the posterior distribution of the parameters for each chain.

To select between models, we used the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). While methods such as the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) only account for model flexibility through the number of free parameters, DIC takes into account the flexibility of the entire functional form of a model (for a detailed explanation, see Evans, Howard, et al., 2017; Myung, 2000). We calculated DIC values for each individual participant, using:

$$\bar{D} = \frac{1}{S}\sum_{s=1}^{S} log[p(y|\boldsymbol{\theta}_s)]$$

$$P_D = max[log[p(y|\boldsymbol{\theta})]] - \bar{D}$$

$$DIC = -2(\bar{D} - P_D)$$

where y are the data for one participant, $\boldsymbol{\theta}$ are the estimated parameters for that participant, and s indexes posterior samples. Some recent studies have criticized DIC, suggesting the Watanabe-Akaike Information Criterion (WAIC) as a superior alternative (Piironen & Vehtari, 2017; Vehtari, Gelman, & Gabry, 2017). However, those studies defined DIC using the posterior mean as the point estimate in the calculation. In contrast, we calculate DIC using the minimum deviance of the posterior distribution as the point estimate (this was also recommended by Spiegelhalter et al., 2002). Recent research has shown that this definition provides near identical results to WAIC in the context of EAMs (Evans, 2019a). Even though the group-level parameters are not directly included in the DIC calculations, the hierarchical estimation indirectly influences the DIC calculation through the constraints they impose on the individual-level parameters ("shrinkage"). We used the DIC values to calculate the "weight" that each model received for each participant, which was done by first transforming the DIC values for each model to the likelihood scale (i.e.,

$exp(\frac{x}{-2})$ from the deviance scale), and then dividing the likelihood for each model by the summed likelihood of both models.

To supplement our formal model comparisons, we visually assessed each model's goodness-of-fit to the joint distribution over response time and accuracy, and also inspected the estimated decision threshold functions. The estimated threshold functions illustrate the size of the collapsing threshold effect in a way that the formal model comparison does not – DIC assesses something closer to statistical reliability than effect size. We assessed goodness-of-fit through quantile probability (Q-P) plots, to ensure that the models provided an adequate description of the data. This is important, for example, to ensure that DIC was not selecting the best of two very poor models. Q-P plots are compact and highly informative, but they can be difficult to read, so we also display how well each model accounted for some standard summary statistics: decision accuracy, and the mean, variance, and skew of response times. These analyses also allow for some level of model-free assessment, as a reduced skew in the response time distributions has been suggested to be a behavioural signature of collapsing thresholds (Hawkins, Forstmann, et al., 2015; Hawkins, Wagenmakers, et al., 2015; Evans & Hawkins, 2019).

Results

Figure 2 summarizes the results of our study, with the rows representing the different experiments. The upper row of Figure 2 displays the results for Experiment 1, with the estimated thresholds in the left column, the quantile probability (Q-P) goodness-of-fit plots in the middle column, and the DIC weights in the right column. The DIC weights show that, for Experiment 1, data from the majority of participants were much better described by the model with fixed thresholds than the model with collapsing thresholds ($p = 0.008$ for a Wilcoxon signed rank test on the DIC difference values). Around 20% of the participants were best described by the collapsing bounds model, and for about another 15% the

analyses were ambiguous. Overall, fixed thresholds – where the quantity of accumulated information required to trigger a decision does not depend on the duration of the decision – was the dominant decision strategy in this reward-rate optimization paradigm.

The Q-P plots (see Donkin & Brown, 2018 for an introduction to Q-P plots) show how well each model accounts for the data by comparing the observed data against posterior predictive data generated from each model. Each dot displays a different response time quantile for one of the four different experimental conditions (coherence), with dots falling above .5 on the x-axis indicating correct responses and dots falling below .5 indicating error responses; for the 0% coherence condition performance was at chance, meaning that there are two dots almost exactly on top of one another for the response time quantiles at .5 on the x-axis. These are response proportions displayed on the x-axis, meaning that dots toward the right of each Q-P panel have a greater proportion of responses associated with them, and therefore carry more weight in the likelihood. The y-axis displays the response time quantiles, with dots that are higher in a column of vertically aligned dots being later response time quantiles. The Q-P plots support the general descriptive adequacy for both models. The Q-P plots also support the conclusions of the model selection, with the fixed thresholds diffusion model providing a better prediction of the majority of quantiles across all conditions, especially in the case of the slower (e.g., .9) quantiles.

Figure 3 compares the summary statistics calculated from data against the predictions of each model, with the first point on the x-axis of each panel providing the results for Experiment 1. Although the collapsing thresholds provides a slightly better account of the decision accuracy, the fixed thresholds model provides a better account of the mean, variance, and skew in responses times for both correct and error responses. This agrees with the model selection results from DIC, that fixed thresholds provide the best account of these data. In addition, the skew in response time for both correct and error responses

appears to be relatively large, suggesting that the data show a pronounced right tail, and therefore, provide model-free evidence in favor of fixed thresholds.

Finally, the estimated decision thresholds of each model are shown as a function of elapsed decision time in the left column of Figure 2. The thresholds from the collapsing thresholds model are much different from the fixed thresholds model. By the 2 second mark, after which responses become sparse, the thresholds have collapsed markedly, though still remain relatively far apart from one another. These group-level estimated thresholds suggest that those participants who showed strong evidence for the collapsing thresholds model demonstrated a large degree of collapse in their thresholds.

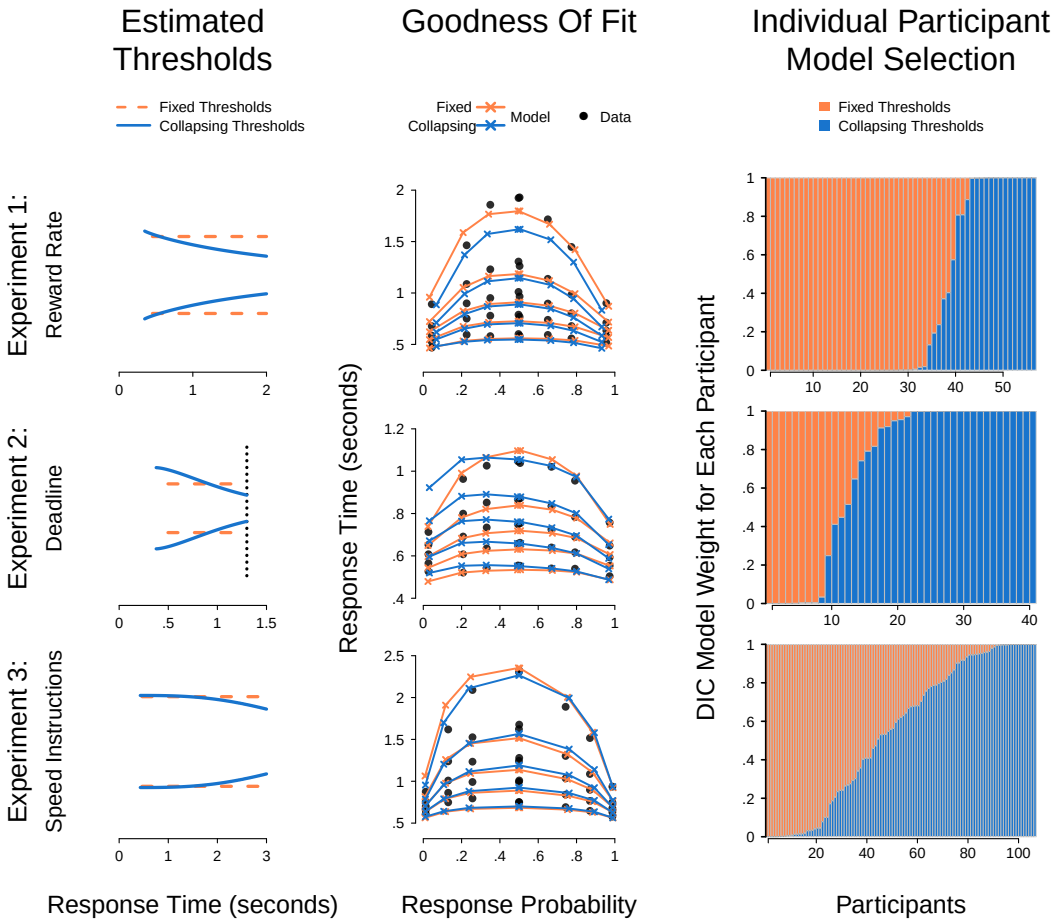

Figure 2. Results from Experiments 1-3 (rows) for each of the three analyses (columns). The left column shows the estimated thresholds for the fixed and collapsing threshold models as a function of predicted response time, the middle column shows quantile-probability (Q-P) plots for each model, and the right column shows the DIC weights for each participant. In all plots, the fixed thresholds model is displayed in orange, and the collapsing thresholds model is displayed in blue. The estimated thresholds for each model were calculated using the median of the group-level mean posterior distributions, and the offset from 0 takes into account the predicted time required for non-decision-related processes. The dashed vertical in Experiment 2 (middle row) represents the response deadline imposed on participants. The Q-P plots were generated by taking the .1, .3, .5 (i.e., median), .7 and .9 response time quantiles for each participant for each response and condition combination. This was done for both the empirical and model predicted data, and then the quantiles were averaged over participants. 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.

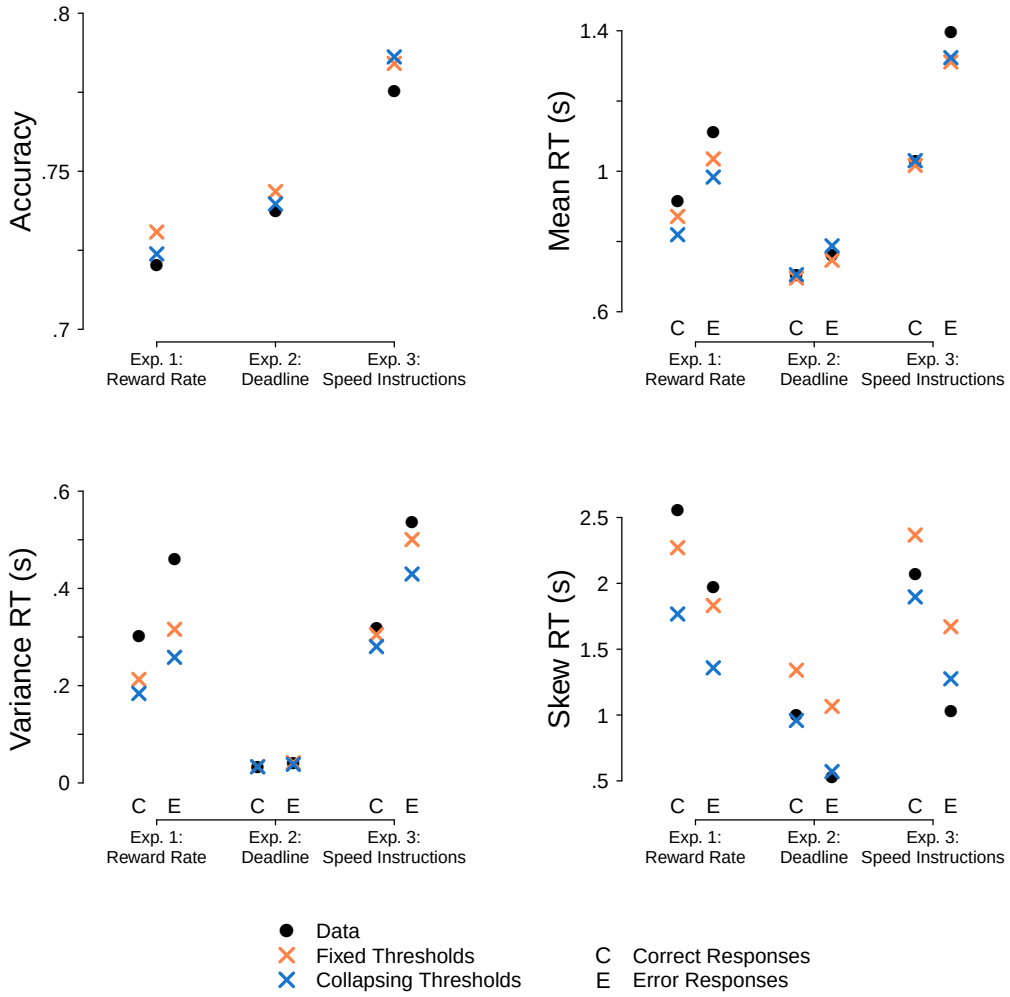

Figure 3. Results from Experiments 1-3 (x-axis) for each of the four summary statistics analyses (different panels; summary statistics plotted as the y-axis). Black dots provide the values for the empirical data, orange crosses provide the predictions for the fixed thresholds model, and blue crosses provide the predictions for the collapsing thresholds model. The C and E on the x-axis provide the summary statistics for the correct and error response time distributions, respectively.

Experiment 2

Experiment 2 investigated one specific method for emphasizing decision speed: a decision deadline.

Method

Participants. Seventy-one undergraduate students from the University of Newcastle completed the experiment online, which was approved by the University of Newcastle Human Research Ethics Committee, and were reimbursed with course credit. Before data collection, we defined an accuracy criterion of 90% for the easiest condition of the task and excluded participants who did not meet this mark (though it should be noted that we did not formally preregister this exclusion criterion). This criterion was more stringent than in Experiment 1 due to the different experimental procedure and easier decisions. We also excluded data from participants who failed to respond before the deadline in more than 10% of their trials. These two exclusions together eliminated data from 30 participants. In the supplementary materials we describe extra analyses of these excluded participants. Those analyses demonstrate that their exclusion would not have changed the overall pattern of results.

Procedure. The experiment was identical to Experiment 1, except that we used a response deadline of 1,300ms, after which the trial would terminate and no response was allowed. In addition, no feedback on performance was provided after any block, so participants were not expected to continue adjusting their thresholds throughout the task. We also used a fixed number of trials, with 10 blocks of 40 trials. Before analysis, the first block of trials was removed to allow for familiarization with the task (see the supplementary materials for analysis showing that performance was approximately stationary after the first block). We removed responses that were faster than 250ms or slower than the

1,300ms deadline, resulting in the removal of less than 3% of the trials from the 41 analyzed participants. We defined the fixed and collapsing thresholds model in the same way as Experiment 1, and all parameter estimation and model selection methods were identical, too.

Results

The middle row of Figure 2 displays the results for Experiment 2, with the estimated thresholds in the left column, the quantile probability (Q-P) goodness-of-fit plots in the middle column, and the DIC weights in the right column. The DIC weights show that data from the majority of participants were best described by the collapsing thresholds model ($p = 0.002$ for a Wilcoxon signed rank test on the DIC difference values), with a few participants showing more ambiguous preferences, and even fewer participants showing a strong preference for the fixed thresholds model. This implies that collapsing thresholds is the dominant decision strategy in this deadline paradigm. Supporting this conclusion, the Q-P plots show that both models account quite well for the data, but that the collapsing thresholds model provides a better description of some parts of the response time distributions; particularly the quantiles associated with the highest probability responses, on the right hand side of each plot. These conditions have the most data in them. In those conditions, the fixed thresholds model underestimates the response time of the fastest quantiles, and overestimates the response time of the slowest quantiles. This corresponds to an over-prediction of the skew of the response time distribution.

Figure 3 displays the summary statistics in the data and the predictions of each model, with the second point on the x-axis of each panel providing the results for Experiment 2. Although both models provide a similar account of the mean and variance in response time for both correct and error responses, the collapsing thresholds model captures the response accuracy slightly better, and provides a better account of the skew in

response time, further supporting the previous conclusions of collapsing thresholds providing the best account of these data. In addition, the skew in response time is smaller than for Experiment 1. This is consistent with a shortening of the right tail of the RT distribution, providing model-free evidence in favour of collapsing thresholds.

The threshold plots show that the decision process is estimated to begin earlier by the collapsing thresholds model than by the fixed thresholds model. For the early stages of the decision process, the thresholds for the collapsing thresholds model are much higher than that of the fixed thresholds model. However, the thresholds for the collapsing thresholds model decrease rapidly as time passes, becoming equal with the fixed thresholds model by about 1s. Lastly, the thresholds of the collapsing thresholds model almost converge by the deadline (1.3s). These empirical trends within the estimated thresholds are in line with the theoretical reasoning for why collapsing thresholds are adaptive in deadline paradigms, as the high initial thresholds allow for high accuracy on easy trials, and the late collapse allows for few trials to be lost to the deadline.

Experiment 3

Experiment 3 investigated another specific method for emphasizing decision speed: speed-emphasis instructions.

Method

Participants. Here we draw on data reported as Experiment 2 of Evans, Rae, et al. (2017). One hundred and fifty-four undergraduate students from the University of Newcastle completed the experiment in lab, and were reimbursed with course credit. We used an identical exclusion criteria to that of Evans, Rae, et al., setting an accuracy criterion of 90% for the easiest condition of the task and excluding participants who did not meet this mark, as well as excluding participants with a mean response time slower than 1.5s,

as these participants failed to comply with the speed-emphasis instructions, which resulted in data from 47 participants being removed. It should be noted that many of these excluded participants favored the collapsing thresholds diffusion model, which would have resulted in a shift in the overall results from relatively ambiguous to some preference for the collapsing thresholds model (see the supplementary materials for the analysis of these excluded participants).

Procedure. The task and procedure were identical to Experiment 1, except for a fixed number of trials (5 blocks of 48 trials) and a time-out after 5s. As with Experiment 2, no feedback on performance was provided after any block, so participants were not expected to continue adjusting their thresholds throughout the task. To induce a speed emphasis, participants were given instructions to respond speedily, both on screen and by the experimenter, before completing the task (see Evans, Rae, et al., 2017 for more details). Before analysis, the first block of trials was removed to allow for practice effects. In contrast to Experiment 2, there was some evidence that mean response time was still changing after this block; see the supplementary materials for analyses. We also removed trials that were faster than 250ms or slower than the 5s timeout, which was approximately 0.3% of trials from the remaining participants. For analyses, we defined the fixed and collapsing thresholds model in the same way as Experiment 1, and used the same parameter estimation and model selection methods.

Results

The bottom row of Figure 2 displays the results for Experiment 3, with the estimated thresholds in the left column, the quantile probability (Q-P) goodness-of-fit plots in the middle column, and the DIC weights in the right column. The DIC weights show that, for the majority of participants, the model discrimination was ambiguous, with both models

doing an almost equally good job of explaining the data ($p = 0.173$ for a Wilcoxon signed rank test on the DIC difference values). In general, there appears to be more overall weight on the collapsing thresholds model, and more participants showing a strong preference for the collapsing thresholds model, suggesting that it provides a slightly better account of these data overall. Also note that the excluded participants tended to show a greater preference for the collapsing thresholds model, meaning that their inclusion would result in an even greater overall preference for the collapsing thresholds model (see supplementary materials for more detail). The Q-P plots support this interpretation, with the collapsing thresholds model providing a better account for those quantiles with the most data (right hand side of the plot), and the fixed thresholds model providing a better account of others.

Figure 3 displays the summary statistics in the data and the predictions of each model, with the last point on the x-axis of each panel providing the results for Experiment 3. Both models provide equally-good accounts of decision accuracy and mean RT; the fixed bounds model provides a much better account of RT variance (particularly for incorrect responses), and the collapsing bounds model provides a much better account of RT skew. This pattern is consistent with the model selection results from the DIC analyses, that the data were ambiguous between these two models. In addition, the skew in response time appears to be somewhat inconsistent across correct and error responses, being fairly large for correct responses – which would suggest evidence for fixed thresholds – and fairly small for error responses – which would suggest evidence for collapsing thresholds. Overall, it is not clear that these data strongly and clearly favour either model.

The threshold plot appears to suggest a relatively weak effect of collapsing thresholds. The thresholds begin at about the same level, with the collapsing threshold beginning to collapse at around the 2 second mark, and showing a minor collapse until about the 3 second mark, where the data become very sparse. However, these thresholds still remain

very close to the estimated fixed thresholds, and remain very far apart from one another. These estimates further support the conclusions of general ambiguity, but potentially a slight preference for the collapsing thresholds model.

General Discussion

Our study compared the fixed and collapsing thresholds accounts of decision-making in paradigms where collapsing thresholds provide theoretical advantages. Previous quantitative comparisons between these models have focused on the typical paradigms used in studies of rapid decision-making, and have generally found an advantage for the fixed thresholds diffusion model (Hawkins, Forstmann, et al., 2015). We investigated three paradigms in which adopting collapsing thresholds can be advantageous to decision-makers: when the goal is optimizing reward rate, when the task contains an explicit deadline, and when the task requires speeded performance. Previous findings suggest that humans do not adopt collapsing thresholds by default, and our findings imply that they also do not adopt collapsing thresholds when instructed to maximize their reward rates. Our Experiments 2 and 3 suggest that most people adopt collapsing thresholds when faced with decision urgency, be it through instructions or an explicit decision deadline. However, the resulting collapse was much greater when participants were given an explicit decision deadline (Exp 2), rather than verbal instructions (Exp 3; though again note that for this experiment, the inclusion of the excluded participants would have made the overall trend more strongly in favor of collapsing thresholds).

A motivation for collapsing threshold theories has been normative accounts that show collapsing thresholds to be optimal for maximizing reward rate when drift rate differs unpredictably between trials (Drugowitsch et al., 2012; Thura et al., 2012). However, it has not been established whether humans attempt to maximize their reward rate by adopting

collapsing thresholds in such circumstances, and our Experiment 1 suggests that they do not. This result is consistent with previous work on reward rate optimality which has demonstrated that humans fail to adopt optimal decision-making policies in many ways (Evans & Brown, 2017; Starns & Ratcliff, 2012, though see Evans, Bennett, & Brown, 2018 for a task-design explanation). Almost every participant (55 out of 57) in our Experiment 1 had a reward rate that was lower than the best possible reward rate under a single fixed threshold (see the supplementary materials for more details). Another advantage of collapsing thresholds models is that they help decision-makers meet demands for urgent decisions, without making many errors. Our Experiments 2 and 3 suggest that humans adopt collapsing thresholds when faced with urgency stress in the form of an explicit looming deadline, and sometimes will adopt collapsing thresholds when told to perform the task quickly. Overall, our findings suggest that the previous normative motivations for collapsing thresholds are not consistent with the decision strategies that humans adopt in situations that encourage the optimization of reward rate, but are consistent with the decision strategies that humans adopt in situations that encourage urgency.

It has been suggested that the differences between fixed and collapsing thresholds models is most prevalent in the slow tails of the response time distribution (Hawkins, Wagenmakers, et al., 2015), but our analyses hint that the comparison of these models might be even more sensitive to the tails of the distribution than previously thought. Previous applications of the fixed thresholds diffusion model have usually included a contamination process (Ratcliff & Tuerlinckx, 2002), which allows for a small proportion of the data to have arisen from a process unrelated to the decision task (e.g., due to distraction or inattention). We analyzed both the collapsing and fixed thresholds diffusion models with and without a contamination process, though we only discuss the versions that included the contamination process, as they generally provided a better account of the data, and have

better theoretical motivations (i.e., participants are surely distracted on some trials). In Experiments 1 and 2, the overall preference (fixed thresholds and collapsing thresholds, respectively) was unchanged regardless of whether or not the contamination process was included. However, in Experiment 3, when comparing the fixed and collapsing bounds models *without* a contamination process, we found evidence in favour of the fixed thresholds model. This is opposite of what was observed in when the contamination process was included (i.e., in the text above), where the evidence was moderately in favour of the collapsing bounds model. This was not a major problem for analysis of Experiment 3, as when including both contamination and non-contamination models in the comparison the pattern of results remained the same (see the supplementary details). We believe that this inconsistency suggests that the contamination process, commonly considered an "auxiliary" assumption, may be more important in the collapsing and fixed thresholds debate than has been understood. A corollary is that some of the ability to differentiate between the models may hinge on the shape of the slow tails of the response time distributions, which can be defined by just a small number of long response times. For these responses, the quick cutoff of the collapsing thresholds model results in an extremely poor likelihood, which penalizes the model, unless these response times are assumed to be due to another process (contamination). Future research will be required to understand these effects, and investigate the validity of the associated assumptions.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723.

Balci, F., Simen, P., Niyogi, R., Saxe, A., Hughes, J. A., Holmes, P., & Cohen, J. D. (2011). Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Attention, Perception, & Psychophysics*, *73*(2), 640–657.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Brown, S. D., Marley, A., Donkin, C., & Heathcote, A. (2008). An integrated model of choices and response times in absolute identification. *Psychological review*, *115*(2), 396.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: the urgency-gating model. *The Journal of Neuroscience*, *29*(37), 11560–11571.

Ditterich, J. (2006a). Evidence for time-variant decision making. *European Journal of Neuroscience*, *24*(12), 3628–3641.

Ditterich, J. (2006b). Stochastic models of decisions about motion direction: behavior and physiology. *Neural Networks*, *19*(8), 981–1012.

Donkin, C., & Brown, S. D. (2018). Response times and decision-making. In E.-J. Wagenmakers (Ed.), *Stevens' handbook of mathematical psychology, volume 5: Methodology.* Wiley.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, *32*(11), 3612–3628.

Evans, N. J. (2019a). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychonomic Bulletin & Review*, 1–29.

Evans, N. J. (2019b). 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. (2018). Bayes factors for the linear ballistic accumulator model of decision-making. *Behavior research methods*, *50*(2), 589–603.

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

Evans, N. J., & Hawkins, G. E. (2019). When humans behave like monkeys: Feedback delays and extensive practice increase the efficiency of speeded decisions. *Cognition*, *184*, 11–18.

Evans, N. J., Hawkins, G. E., Boehm, U., Wagenmakers, E.-J., & Brown, S. D. (2017). The computations that support simple decision-making: A comparison between the diffusion and urgency-gating models. *Scientific reports*, *7*, 16433.

Evans, N. J., Howard, Z. L., Heathcote, A., & Brown, S. D. (2017). Model flexibility analysis does not measure the persuasiveness of a fit. *Psychological review*, *124*(3), 339.

Evans, N. J., Rae, B., Bushmakin, M., Rubin, M., & Brown, S. D. (2017). Need for closure is associated with urgency in perceptual decision-making. *Memory & Cognition*, 1–13.

Forstmann, B. U., Dutilh, G., Brown, S. D., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre–SMA facilitate decision–making under time pressure. *Proceedings of the National Academy of Science*, *105*, 17538–17542.

Forstmann, B. U., Tittgemeyer, M., Wagenmakers, E.-J., Derrfuss, J., Imperati, D., & Brown, S. (2011). The speed-accuracy tradeoff in the elderly brain: a structural model-based approach. *The Journal of Neuroscience*, *31*(47), 17242–17249.

Frazier, P. I., & Yu, A. J. (2007). Sequential hypothesis testing under stochastic deadlines. In *Nips* (pp. 465–472).

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*(3), 389.

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic bulletin & review*, *19*(2), 339–348.

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.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive science*, *38*(4), 701–735.

Hawkins, G. E., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Discriminating evidence accumulation from urgency signals in speeded decision making. *Journal of neurophysiology*, *114*(1), 40–47.

Ho, T. C., Yang, G., Wu, J., Cassey, P., Brown, S. D., Hoang, N., ... others (2014). Functional connectivity of negative emotional processing in adolescent depression. *Journal of affective disorders*, *155*, 65–74.

Holmes, W. R. (2015). A practical guide to the probability density approximation (pda) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68*, 13–24.

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.

Matzke, D., Dolan, C. V., Logan, G. D., Brown, S. D., & Wagenmakers, E.-J. (2013). Bayesian parametric estimation of stop-signal reaction time distributions. *Journal of Experimental Psychology: General*, *142*(4), 1047.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*(1), 190–204.

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.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of vision*, *5*(5), 1–1.

Piironen, J., & Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

Pilly, P. K., & Seitz, A. R. (2009). What a difference a parameter makes: A psychophysical comparison of random dot motion algorithms. *Vision Research*, *49*(13), 1599–1612.

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.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review*, *85*(2), 59.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences*, *20*(4), 260–281.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic bulletin & review*, *9*(3), 438–481.

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of neuroscience*, *22*(21), 9475–9489.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791–806.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1865.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures

of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and aging*, *25*(2), 377.

Starns, J. J., & Ratcliff, R. (2012). Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. *Psychonomic bulletin & review*, *19*(1), 139–145.

Stone, M. (1960). Models for choice–reaction time. *Psychometrika*, *25*, 251–260.

Thura, D., Beauregard-Racine, J., Fradet, C.-W., & Cisek, P. (2012). Decision making by urgency gating: theory and experimental support. *Journal of Neurophysiology*, *108*(11), 2912–2930.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological methods*, *18*(3), 368.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, *27*(5), 1413–1432.

Voskuilen, C., Ratcliff, R., & Smith, P. L. (2016). Comparing fixed and collapsing boundary versions of the diffusion model. *Journal of Mathematical Psychology*, *73*, 59–79.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 326–339.

Winkel, J., Keuken, M. C., van Maanen, L., Wagenmakers, E.-J., & Forstmann, B. U. (2014). Early evidence affects later decisions: Why evidence accumulation is required to explain response time data. *Psychonomic Bulletin & Review*, *21*(3), 777–784.