Speed-Accuracy Tradeoffs in Decision Making: Perception Shifts and Goal Activation Bias Decision Thresholds

Jeffrey S. Larson[1] & Guy E. Hawkins[2]

1: Department of Marketing and Global Supply Chain, Marriott School of Business, Brigham Young University, Provo, UT 84606, jeff_larson@byu.edu

2: School of Psychology, University of Newcastle, Newcastle, NSW Australia 2308 guy.hawkins@newcastle.edu.au

Abstract

A fundamental aspect of decision making is the speed-accuracy tradeoff (SAT): slower decisions tend to be more accurate, but since time is a scarce resource people prefer to conclude decisions more quickly. The current research adds to the SAT literature by documenting two previously unrecognized influences on the SAT: perception shifts and goal activation. Decision makers' perceptions of what constitutes a *fast* or a *slow* decision, and what constitutes an *accurate* or *inaccurate* decision, are based on prior experience, and these perceptions influence decision speed. Similarly, previous experience in a decision context associates the context with a particular decision goal. Thus, in later decisions the decision context will activate this goal, and thereby influence decision speed. Both of these mechanisms contribute to a specific decision bias: decision speeds are biased toward original decision speeds in a decision context. Four experiments provide evidence for the bias and the two contributing mechanisms.

Keywords: Speed-accuracy tradeoffs, decision threshold, goals, perceptions.

Introduction

The speed-accuracy tradeoff (SAT) is a crucial property of all decision making. Slower and more effortful decisions increase the likelihood of an accurate decision, but they consume more of a valuable and limited resource (time). Faster decisions conserve time but are less likely to result in selection of the most desirable outcome. Research from various traditions has made important contributions to our knowledge of how these tradeoffs are made in practice. By one research tradition, the SAT is mediated by the selection of one of many possible decision strategies that vary in speed and accuracy (Payne, Bettman, & Johnson, 1988; Rieskamp & Otto, 2006). An alternate research tradition posits that the SAT is governed by a decision threshold—decision makers process evidence about decision alternatives until the evidence for one of the alternatives reaches a threshold, at which time that alternative is selected (for reviews, see Forstmann, Ratcliff, & Wagenmakers, 2016; Ratcliff, Smith, Brown, & McKoon, 2016; Voss, Nagler, & Lerche, 2013). The setting of this threshold is the primary factor that determines the SAT. A high threshold means more evidence must be accumulated for an option, and thus resulting decisions are slow but more accurate. A low threshold requires less evidence, which results in faster decisions that are less accurate. This latter tradition has dominated SAT research for decades and has greatly advanced our knowledge of the SAT. In the current research, we adopt this tradition in our analysis of the SAT but also account for the possible role of decision strategies.

The "height" of the threshold—which represents the quantity of evidence required to commit to a choice—is the primary mechanism by which decision makers trade off speed and accuracy, so understanding how decision makers set this threshold is fundamental to our understanding of the SAT. Extant research has examined the process by which decision makers set or modulate the threshold in a decision task (Gold & Shadlen, 2007; Ratcliff, Van Zandt, & McKoon, 1999; Simen, Cohen, & Holmes, 2006), the influences of decision task properties on threshold levels (for review, see Heitz, 2014), how thresholds relate to reward rates (Balci et al., 2011; Bogacz, Hu, Holmes, & Cohen, 2010; Evans & Brown, 2017; Simen et al., 2009; Starns & Ratcliff, 2012) or other goal maximizing strategies (e.g., Hawkins,

Brown, Steyvers, & Wagenmakers, 2012a, 2012b; Zacksenhouse, Bogacz, & Holmes, 2010), how thresholds might vary as a function of decision time (e.g., Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015; Palestro, Weichart, Sederberg, & Turner, 2018), and even the neural circuitry governing the SAT (e.g., Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Forstmann et al., 2010, 2008; Heitz & Schall, 2012; Van Vugt, Simen, Nystrom, Holmes, & Cohen, 2014). All of this prior research is based on the assumption that decision makers accurately assess the speed and accuracy of their decisions. In the current research, we show that these assessments are readily biased by prior experience. A person's perceptions of *fast* and *slow* and *accurate* and *inaccurate* are subject to perceptual biases that can greatly influence thresholds.

We also document a second influence on thresholds that is hitherto unrecognized in the SAT literature, and that is the influence of decision goals. The competing decision goals to make a decision quickly and to make a decision accurately form the basis of the SAT, and therefore influence all decisions. An immense literature stream has examined and continues to examine goal activation, striving and fulfillment, but no prior research has investigated how the known properties of goal activation influence decision thresholds. The current research demonstrates that decision goals can be activated by the decision context and this activation can bias decision makers' thresholds.

Both of these previously uninvestigated influences on decision thresholds, perception shifts and goal activation, influence thresholds in the same way—decision makers underadjust thresholds in the face of decision task changes that require them to increase or decrease their threshold. That is, a decision maker who makes fast decisions in one decision context will continue making fast decisions when a change in the context necessitates slower decisions, and a decision maker who makes slow decisions in one decision context will continue making slow decisions when a change in the context necessitates faster decisions. In essence, we show that thresholds are "sticky". To our knowledge, such an effect has not been demonstrated in any of the prior research on the SAT, threshold setting, or threshold modulation.

In summary, this article makes a threefold contribution to the SAT literature. First, it provides evidence for a previously unrecognized phenomenon: decision thresholds are biased toward previous threshold levels when a change in the decision task necessitates an adjustment to the threshold level. This phenomenon has two underlying causes: perception shifts and goal activation. The second and third contributions of this article are the demonstration of the roles of these mechanisms in causing biased thresholds. But these two contributions extend beyond causing the demonstrated bias in thresholds; the fact that decision thresholds in repeated decisions are affected by perceptions and goals is novel to the SAT literature.

The next section provides an explanation of the modeling framework of the SAT literature and reviews important findings from that literature. Following this literature review, we explain the effect of sticky thresholds and how perception shifts and goal activation both produce this effect. Four experiments provide evidence for the proposed phenomenon and the two causal mechanisms. Data from these studies are analyzed using both conventional statistics and a quantitative SAT model. Finally, we conclude with discussion of the results and suggestions for future research.

Cognitive Models of the Speed-Accuracy Tradeoff

The vast majority of modern cognitive process models of the SAT assume that decisions are made through a process of evidence accumulation. Although there is a large family of evidence accumulation models that differ in minor details (for reviews, see Forstmann et al., 2016; Ratcliff & Smith, 2004; Ratcliff et al., 2016), almost all make the same three basic assumptions (though for exceptions, see, e.g., Hawkins & Heathcote, 2021; Verdonck & Tuerlinckx, 2014). First, the decision maker accumulates evidence in favor of each choice option over time. Second, a response is made once sufficient evidence has accumulated for one choice option over the other/s. Third, there is an offset time for response-related components that occur outside of the choice process, such as the time required to encode the stimulus information and produce a motor response. Taken together, these three assump-

tions form the foundation of evidence accumulation models: decisions are made through a process of gradually accumulating information to a threshold.

Figure 1 provides an illustration of a single decision made using the choice process of the SAT model that we study in this paper—the Linear Ballistic Accumulator (LBA) model (Brown & Heathcote, 2008). In the LBA, the decision maker samples information and accumulates evidence (assumption #1) separately for each response option. The evidence accumulated in favor of each response is illustrated as its height on the y-axis. In the decision illustrated in Figure 1, option A (left) has accumulated more evidence than option B (right). The process continues over time until the evidence for one of the options reaches the response threshold (dotted line), which triggers a response (assumption #2). The predicted response corresponds to the option that first crossed the response threshold (option A in Figure 1). The predicted response time is the time it took for the accumulated evidence to reach the threshold (i.e., x-axis position at the threshold-crossing point) plus an offset time that accounts for components that are peripheral to the decision itself, including encoding the stimulus display and producing a physical response (such as a button press or eye movement—assumption #3; not illustrated in the figure).

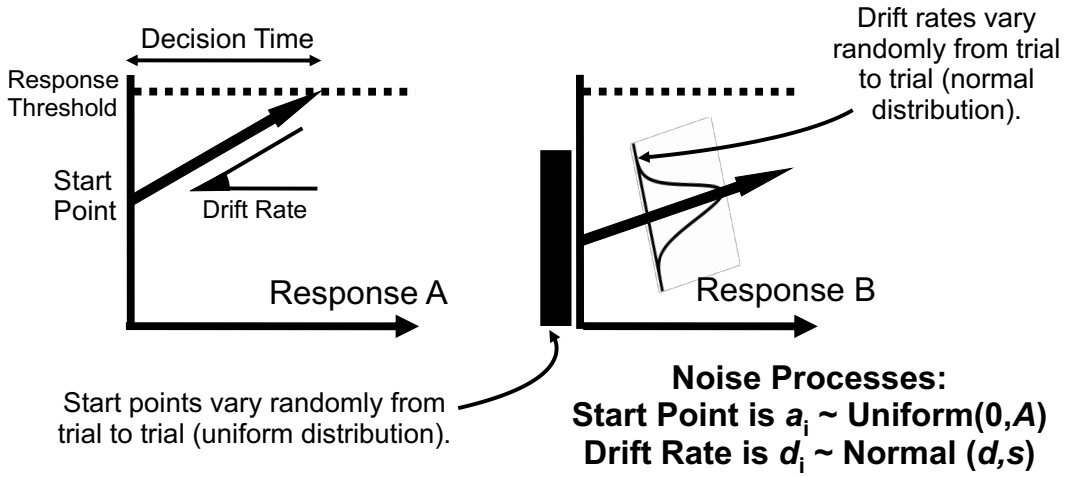

Figure 1. Illustration of the SAT framework in two-alternative forced choice settings.

The parameters of SAT models reflect the latent constructs that are hypothe-

sized to underlie observed decision behavior. For example, the average rate of evidence accumulation—known as the drift rate—indexes the mean speed with which evidence in favor of each choice option is accumulated from the stimulus, where a larger drift rate indicates faster accumulation of evidence for that choice option; this reflects greater utility for that choice option. The magnitude or "height" of the response threshold indicates the level of response caution, where a higher threshold indicates that decision makers require more evidence for an option before they are willing to commit to selecting an option. The time taken for aspects of the observed response time that is not accounted for by the choice process itself—stimulus encoding, executing a physical response to indicate a choice—is known as non-decision time. The value of the cognitive interpretations of the SAT model parameters have been validated through tests of selective influence—a priori hypotheses about the effect of particular experimental manipulations on the latent components of the SAT model—across a range of studies. For example, Voss, Rothermund, and Voss (2004) showed that making the motor component of producing a response more challenging only led to increases in non-decision time. Similarly, Ratcliff and Rouder (1998) showed that manipulating task difficulty led to changes in drift rate (processing speed) but not response threshold (cautiousness), and instructions that emphasized decision speed or accuracy led to changes in response threshold but not drift rate; though others have observed other patterns in data (e.g., Dutilh et al., 2019; Hawkins & Heathcote, 2021; Rae, Heathcote, Donkin, Averell, & Brown, 2014). In addition, past research shows tight correspondence between specific neural structures and parameters of SAT models (Bogacz, Wagenmakers, et al., 2010; Gold & Shadlen, 2007; Heitz & Schall, 2012; Van Vugt et al., 2014).

Years of psychological research have shown that SAT models provide a good account, across a wide array of stimuli, of the tension between decision speed and decision accuracy through the setting of the response threshold (Forstmann et al., 2016; Hawkins et al., 2012a, 2012b; Ratcliff et al., 2016; Roe, Busemeyer, & Townsend, 2001, though see Rae et al. 2014 and Evans 2021 for an alternative account). Since thresholds can be adjusted in response to task demands, such as the instruction to respond fast or accurately, the level of caution

in choice is thought to be under the strategic control of the decision maker.

Setting the Speed-Accuracy Tradeoff in Changing Decision Environments

Models of the SAT explain how, through the setting of a threshold level, a decision maker can trade decision speed for decision accuracy. It has long been known that if the correct decision can be identified objectively and that both the decision difficulty and the reward for correct answers is constant across decisions, then the sequential probability ratio test (Wald & Wolfowitz, 1948) and SAT models (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) can determine the threshold setting that is optimal for maximizing reward. Prior research has documented a behavioral bias in many participants' setting of the threshold. Specifically, they find that participants tend to set their thresholds higher than the level that would maximize financial reward (Balci et al., 2011; Bogacz, Hu, et al., 2010; Evans & Brown, 2017; Simen et al., 2009; Starns & Ratcliff, 2012). In the current research, we are less interested in global optimality of the threshold setting and instead focus on how participants adjust their threshold to changes in the incentive structure of the decision context that necessitate an adjustment to the threshold. The reason for this focus is twofold. First, many real-life decisions—what to wear, what to eat, what to buy—are made repeatedly. Even if the choice context of a repeated decision is not perfectly identical to the previous decision, many aspects of the context are likely preserved from one choice occasion to the next. By studying contextual influences in repeated decisions, we are examining factors that influence a good share of decisions. Second, the mechanisms we investigate come about precisely because of the repeated nature of decisions. Still, the focus on repeated decisions generates novel insights about speed-accuracy tradeoffs generally, and not just the SAT in repeated decisions.

Extant research on repeated choice tasks finds that participants adapt quickly to changes in task conditions (Bogacz et al., 2006; Ratcliff et al., 1999). Importantly, while this research determined that adaptation took place quickly, it did not examine whether the new thresholds were systematically influenced by the prior thresholds. Several researchers

have proposed models for the mechanism by which decision makers set thresholds in new choice tasks (Myung & Busemeyer, 1989; Simen et al., 2006; Vickers & Lee, 1998, 2000), but none of these models proposes that thresholds might be systematically biased in the sense that they become maladaptive in the current decision context (too cautious or not cautious enough). Our research also differs from previous work demonstrating that the bias to give a particular response is systematically adapted as a function of trial history (e.g., Treisman & Williams, 1984). In the current research, we identify a systematic bias, not in the response itself but in the setting of the threshold that determines response caution during repeated choices. Specifically, when conditions in a choice task change in a manner that requires participants to adjust their decision threshold, the resulting adjustments are typically insufficient. In other words, thresholds exhibit a bias toward earlier threshold levels.

The finding that prior experience can bias overall threshold levels, rather than the bias to give one response over another, is an important contribution to the SAT literature for several reasons. Research in the domain of SAT utilizes repeated choices, sometimes on the order of thousands of choices, so to find evidence for biased thresholds would have important implications for the design of SAT experiments. In practical settings, because many decisions are repeated, biased thresholds could be the genesis of persistent problematic behavior. For example, a person may frequently be late because he/she consistently spends too long deciding what to wear, due to high thresholds set in earlier decisions, which upwardly bias the threshold even on days when the person is in a hurry. In contrast, a different person may show up for an important meeting dressed inappropriately because he/she spent too little time deciding what to wear, which resulted from a downward bias in threshold established by early experience. Further, though the bias most readily occurs in a repeated decision environment, the bias may manifest in decisions that share some contextual characteristics. For example, a particular shopping environment may activate a specific threshold for all decisions that occur within the store, even when such decisions are not repeated. But perhaps the most important contribution made by this article is the

demonstration of two influences on decision thresholds that have not been recognized in prior literature.

Perception Shifts

Perception shifts refer to changes in decision makers' beliefs about what it means to make *fast* or *accurate* choices. Extensive research in psychophysics has established that perceptions of speed (and countless other phenomena) are highly dependent on the range of prior stimulus experience (Helson, 1964; Stevens, 2017). Therefore, if a decision maker made all previous decisions in a particular context in 10 seconds or less, then she would perceive a decision in that context lasting 15 seconds to be *slow*. In contrast, if a decision maker had spent at least 20 seconds on all previous decisions in that same context, she would perceive a decision in that context lasting 15 seconds to be *fast*.

Perception shifts lead to biased threshold levels in a straightforward manner. Even when a decision maker appropriately recognizes that a change in the decision context necessitates a faster decision, if she perceives a decision lasting 15 seconds to be *fast*, she is unlikely to make decisions in under 10 seconds when required. Similarly, if another decision maker perceives 15 seconds to be *slow*, she is unlikely to spend a full 25 seconds on a decision when that is required to reach the desired level of accuracy.

Perception shifts in the domain of accuracy can also contribute to biased thresholds. If a decision maker recognizes the need to achieve *high accuracy*, she will perceive an 80% accuracy level to have met this requirement if her prior fast decisions were at chance level accuracy. In contrast, a decision maker who previously made near-perfect decisions would perceive 80% accuracy to be low.

Extant research has not recognized the potential for perception shifts to influence speed-accuracy tradeoffs. Prior accounts of threshold setting propose that decision makers adjust thresholds in response to feedback from each decision (Busemeyer & Rapoport, 1988; Gold & Shadlen, 2007; Myung & Busemeyer, 1989; Simen et al., 2006; Vickers & Lee, 1998, 2000), and this research assumed that decision makers interpreted this feedback accurately.

If decision makers' perceptions of feedback from these decisions can be shifted by prior experience, as we propose, such a phenomenon would be fundamental to the understanding of the SAT.

Goal Activation

In any decision, it is typically assumed that at least two goals are active: (1) make an accurate decision and (2) make the decision quickly. However, the relative strength of these goals helps to determine the SAT. In some decision contexts, a decision maker may consistently set a strong goal to make the decision quickly (and necessarily forego accuracy). In other contexts, a decision maker may consistently set a strong goal to be as accurate as possible (and necessarily take longer on the decision). After repeated choices in a decision environment, the environment itself will activate the original goal. According to (Bargh & Barndollar, 1996, p. 464), "If an individual frequently and consistently chooses the same goal within a given situation, that goal eventually will come to be activated by the features of that situation and will serve to guide behavior, without the individual's consciously intending, choosing, or even being aware of the operation of that goal within the situation." Therefore, if a decision maker consistently sets a strong goal to make a decision quickly in a particular context, the decision environment itself will activate that goal in future decisions, even if the decision maker has explicitly changed her goal. In other words, the decision context becomes a conditioned stimulus and the decision goal the conditioned response. This automatic activation of goals also contributes to biased decision thresholds. To our knowledge, no prior research on SAT has recognized how the properties of goal activation might influence thresholds. Most of the SAT literature could be characterized as an account of competing goals (i.e., balancing the competing demands of speed vs accuracy), but ours is the first to demonstrate how goal activation functions in the SAT.

A third potential mechanism that could explain biased thresholds is decision strategy carryover. In most of the recent research on the SAT, decision makers are assumed to use

the same decision strategy across decisions—that is, a strategy of evidence accumulation—regardless of decision speed. But earlier research has proposed that decision makers adopt different strategies in response to differing needs for speed or accuracy (Payne et al., 1988). If a decision maker adopts a fast decision strategy and continues to apply this same strategy when the context calls for slower decisions, the carryover of this strategy could also contribute to the same bias in the decision threshold. Decision strategy carryover has been demonstrated in prior research (Bröder & Schiffer, 2006; Levav, Reinholtz, & Lin, 2012), and we do not explicitly examine decision strategies here. Instead, we aim to demonstrate the unique contributions of perception shifts and goal activation on thresholds, while controlling for the potential effect of decision strategy carryover.

In four experiments, we provide evidence for biased thresholds and for the two explanatory mechanisms. In Experiment 1, we show evidence for biased thresholds in both directions. We analyze the data using separate analyses of speed and accuracy and compare these results with a simultaneous model of speed and accuracy from the SAT tradition. Further, we provide evidence for the first explanatory mechanism, perception shifts, through reported evaluations of speed and accuracy made by the participants. In Experiment 2, we again demonstrate biased thresholds and provide evidence for the second explanatory mechanism, goal activation. As in Experiment 1, Experiment 2 will show evidence for biased thresholds using both a model from the SAT tradition and separate analyses of speed and accuracy. Experiment 3 provides additional evidence of the influence of goal activation in biasing decision thresholds. Finally, Experiment 4 provides direct evidence of perception shifts in a choice context that enables finer measurement of decision strategies to rule out this alternative mechanism.

Experiment 1

The primary aim of this study is to provide evidence for the proposed bias in the setting of overall threshold levels. A secondary aim is to provide evidence that perception shifts are a plausible cause of this bias.

Method

The 84 participants in this experiment were members of a paid subject pool from a private university in the eastern United States. Participants engaged in a series of unrelated studies in a one-hour study session conducted in an on-campus computer lab in groups of 8 to 12. At the end of the one-hour session, participants received $10 plus their earnings from this study, which averaged $5.15.

Participants engaged in repeated choices in the same decision environment. The task was a two-tier pricing problem in which participants were asked to determine which of three cell phone pricing plans would yield the lowest bill for a given level of usage. They were given the usage for that month, the base price (for usage of 500 or fewer minutes), and the overage rate for three different plans (the additional price for each minute of usage over 500). These values were randomly generated from a uniform distribution in every set (usage between 1 and 1500 minutes; base rate between $29 and $61; overage rate between $.01 and $.15 per minute). An example trial from the experiment is shown in Figure 2. This task is ideal for our purposes because it enables measurement of decision accuracy, since every set contains an objectively correct option (i.e., the option with the lowest price). In addition, this task makes clear to participants from the outset how they can obtain the correct answer, which limits learning effects to the learning that interests us—learning to trade speed and accuracy rather than learning the parameters of a correct decision.

Participants performed the task in four 2-minute phases, preceded by a 45-second training phase to accustom participants to the controls. In each phase, participants received instructions on how they would be paid for their performance. We aimed to manipulate response thresholds by creating two different payment schemes, Fast and Careful. In the Fast payment scheme, participants earned $.25 times their percent correct plus $.04 per correct answer. In the Careful payment scheme, participants earned $1.00 times their percent correct, plus $.01 per correct answer. Participants in the Careful payment scheme received these instructions: “You will now undertake four 2-minute phases of this pricing task. At the end of each phase, you will be paid $1 times your percentage correct, plus

Figure 2. Experiment 1: Example trial.

$.01 for every question you get correct. For example, if you make 25 guesses and are correct on 20 of them, you will earn $1.00 ($.80 for attaining 80% accuracy, and $.20 for 20 correct guesses). Because you are being paid more for accuracy than speed, you should GO SLOWLY." Participants in the Fast condition received comparable instructions: "You will now undertake four 2-minute phases of this pricing task. At the end of each phase, you will be paid $0.25 times your percentage correct, plus $.04 for every question you get correct. For example, if you make 25 guesses and are correct on 20 of them, you will earn $1.00 ($.20 for attaining 80% accuracy, and $.80 for 20 correct guesses). Because you are being paid more for speed than accuracy, you should GO QUICKLY."

In both conditions, after each response, the word "Correct!" or "Wrong!" flashed briefly on the screen. To ensure that the results were not driven by the existence of this explicit accuracy feedback, no accuracy feedback was provided in Experiment 2. After receiving these instructions, participants had two minutes to earn as much money as they could by repeatedly choosing the least expensive option from sets of three cell phone pricing

plans.

The aim of this study was to test for biased thresholds caused by prior thresholds set in an earlier decision context. To do this, we examined choice behavior after a change in the incentive structure. Half of the participants faced a changed incentive in the third phase of choices, while the other half of the participants did not experience a change in the incentive condition, so as to provide a benchmark for comparison. This yielded four conditions: 1) Fast-Fast (F-F), 2) Fast-Careful (F-C), 3) Careful-Careful (C-C), and 4) Careful-Fast (C-F). Participants in the F-C condition, for example, made choices under the Fast incentives for the first two 2-minute choice phases, then made choices under the Careful incentives for the last two 2-minute choice phases.

After every phase of choices, participants rated their perceptions of both their speed and their accuracy on a 1 to 7 semantic differential scale, anchored by "Very Slow" and "Very Fast" or "Not at all Accurate" and "Very Accurate", and were then informed of their accuracy and earnings in the previous phase (number of decisions made, number of correct decisions, percent correct, and monetary earnings).

Results

We present the results of this experiment using two different analysis approaches. We begin with a conventional statistical analysis that independently assesses choice speed with an ANOVA model and choice accuracy with logistic regression. We follow the conventional analysis approach with an SAT model (the LBA model) applied to the data. The SAT model analyzes the joint distribution of choices and response times and thus circumvents the need to conduct independent analyses of speed and accuracy data to assess the SAT. This means it provides a more direct test of the central hypothesis while also addressing some alternate explanations that the conventional analyses cannot. Importantly, the SAT model is assessed using the complete distribution of choices and response times from every participant, which provides a much more sensitive—and hence powerful—assessment of the data than the summary statistics (i.e., cell means) used in the conventional analysis. How-

ever, this sensitivity comes with the potential downside that SAT models can be influenced by outliers, and therefore the data require some cleaning prior to analysis. The conventional analysis approach can be done without this data cleaning. The combined results provide converging evidence and strongly corroborate our main findings.

Conventional Statistical Analysis. We analyzed response time data with an ANOVA using the mixed procedure in SAS. We employed a repeated measures statement to allow the model to account for the correlated nature of within-participant responses, so that data from participants making more responses than average are not over-weighted. (Because each phase lasted two minutes, and participants chose their speed, the number of responses from each participant varies.) We analyzed accuracy data with a logistic regression using the nlmixed procedure, which allows for fitting of a random-effects model on subjects' intercepts. Additionally, response times were transformed by a natural logarithm because of right skew in the distribution of response times, consistent with recommended procedures for timing data (Kalbfleisch & Prentice, 1980).

First, we examine the evidence for biased thresholds, which should manifest in phase 3 of the experiment as: (1) faster choice times for F-C participants than C-C participants; (2) lower accuracy for F-C participants than C-C participants; (3) slower choice times for C-F participants than F-F participants; (4) higher accuracy for C-F participants than F-F participants. This same pattern is also expected to appear in phase 4. Table 1 shows the mean response time and accuracy, along with estimated standard errors, for each condition across the 4 phases. (Analysis of both response time and accuracy was conducted on transformed data—the natural log transformation in the case of response time and the logit transformation in the case of accuracy. Table 1 reports the means back-transformed to the original scale. Standard errors were back-transformed using Taylor expansion.) We report means on the original scale, and display the distribution of mean response times and accuracy rates in Figure A1 (Appendix A).

Consistent with the hypothesized bias in thresholds, F-C participants made choices more quickly ($M_{F\text{-}C}$ = 3.9 s) than C-C participants ($M_{C\text{-}C}$ = 4.6 s) in phase 3, ($F(1, 320)$

Table 1
Experiment 1: Mean response time and accuracy by condition and phase.

Average Response Time in seconds (standard error)

	Phase 1	Phase 2	Phase 3	Phase 4
F-F	3.1 (.08)	2.7 (.07)	2.4 (.05)	2.2 (.05)
F-C	3.0 (.08)	2.5 (.06)	3.9 (.12)	3.7 (.11)
C-C	6.0 (.24)	5.2 (.18)	4.6 (.16)	4.6 (.16)
C-F	6.0 (.22)	4.8 (.16)	3.0 (.08)	2.9 (.07)

Average Accuracy in percent (standard error)

	Phase 1	Phase 2	Phase 3	Phase 4
F-F	72.6 (5.1)	72.2 (5.1)	69.5 (5.6)	70.1 (5.5)
F-C	76.5 (4.4)	71.5 (5.4)	78.5 (4.1)	87.9 (2.3)
C-C	92.1 (1.6)	85.2 (2.8)	90.6 (1.8)	85.9 (2.7)
C-F	93.0 (1.0)	90.3 (1.1)	77.9 (2.0)	75.4 (3.2)

Fast Condition

= 16.1, $p < .01$). A similar trend was seen in the accuracy data, as F-C participants demonstrated significantly lower accuracy ($M_{F\text{-}C}$ = 78.5%) than C-C participants ($M_{C\text{-}C}$ = 90.6%; t(80) = 2.8, $p < .01$). Evidence for sticky thresholds was also manifest in the opposite direction, as C-F participants exhibited slower choices ($M_{C\text{-}F}$ = 3.0 s) than F-F participants ($M_{F\text{-}F}$ = 2.4 s; F(1,320) = 55.6, $p < .01$), and accuracy rates followed the expected direction, with C-F participants exhibiting higher accuracy ($M_{C\text{-}F}$ = 77.9%) than F-F participants ($M_{F\text{-}F}$ = 69.5%) though this effect did not reach significance (t(80) = 1.6, $p = .10$).

In phase 4, choice speeds remained significantly slower for C-F compared to F-F participants (F(1, 320) = 69.0, $p < .01$) and F-C participants were significantly faster than C-C participants (F(1, 320) = 126.9, $p < .01$). Accuracy rates generally followed the same directional trends for the contrast between C-F and F-F participants but did not reach significance in phase 4, while accuracy for F-C and C-C was approximately equal across participants (both ps > .05). Differences in response speed manifest more reliably than differences in accuracy due to the continuous versus binary nature of response speed versus accuracy, respectively.

Experiment 1 also provides evidence in favor of perception shifts as an explanatory mechanism underlying the threshold differences. After every phase of decisions, participants rated the perceived speed and accuracy of their choices. However, ratings from F-C and C-C participants cannot be compared directly, because the ratings provided by participants reflect two different influences: (1) differences in *perceived* speed and (2) differences in *actual* speed. In order to examine how ratings reflect a difference in *perceived* speed, we must adjust the ratings to the extent that they reflect a difference in actual speed. To do this, we modeled self-rated speed as a function of condition and phase, but we included the number of responses made as a covariate. Including this measure of speed as a covariate removes the influence of *actual* speed so that differences in self-rated speed reflect only differences in *perceived* speed.

In phase 3, F-C participants reported a significantly slower perceived speed ($M_{F\text{-}C} = 2.8$) than C-C participants ($M_{C\text{-}C} = 4.8$; $t(80) = 6.3$, $p < .01$). This perception continued to phase 4 ($M_{F\text{-}C} = 2.4$ vs. $M_{C\text{-}C} = 4.5$; $t(80) = 6.5$, $p < .01$). In the opposite direction, C-F participants ($M_{C\text{-}F} = 5.8$) reported significantly faster perceived speed than F-F participants ($M_{F\text{-}F} = 5.2$; $t(80) = 2.1$, $p = .04$). This effect also continued to phase 4 ($M_{C\text{-}F} = 5.8$ vs. $M_{F\text{-}F} = 5.0$; $t(80) = 2.8$, $p < .01$).

We also measured participants' perceptions of their accuracy. As with speed perceptions, the model of self-rated accuracy included the true accuracy as a covariate so that any differences in reported accuracy reflect only a difference in *perceived* accuracy. In phase 3, C-F participants ($M_{C\text{-}F} = 4.2$) perceived themselves to have a lower accuracy than F-F condition participants ($M_{F\text{-}F} = 5.0$; $t(80) = 3.0$, $p < .01$). In the opposite direction, F-C participants ($M_{F\text{-}C} = 4.7$) reported a higher perceived accuracy than C-C participants ($M_{F\text{-}C} = 4.2$), but this difference was not significant ($t(80) = 1.4$, $p = .17$).

Because the decision task had an objectively correct answer and the monetary incentives were known, we can examine performance relative to an optimality benchmark. Performance is optimal if an individual sets his/her threshold such that the expected monetary reward is maximized. Whether sticky thresholds lead to less-optimal performance

(compared to participants who do not face a change in the incentives) is not a key issue in the current investigation, but comparing performance levels of participants who face a change in incentives versus those who do not provides some insight. The results of Experiment 1 indicate that threshold anchoring may lead to lower performance. In both phases 3 and 4, C-F participants earned less (M_{Phase3} = \$1.03, M_{Phase4} = \$1.06) than F-F participants (M_{Phase3} = \$1.14, M_{Phase4} = \$1.24; F(1, 320) = 3.83, p = .05; F(1, 320) = 11.04, $p <$ .01). In phase 3, F-C participants earned less ($M_{F\text{-}C}$ = \$.96) than C-C participants ($M_{C\text{-}C}$ = \$1.04), but this difference was not significant (F(1, 320) = 2.39, p = .12). Due to the anomalously high accuracy of F-C participants in phase 4, they did not earn less than C-C participants in that phase.

Overall, these results are consistent with past research. Past research has repeatedly shown that people tend to set their threshold too high relative to the threshold that would maximize rewards (Balci et al., 2011; Bogacz, Hu, et al., 2010; Evans & Brown, 2017; Simen et al., 2009; Starns & Ratcliff, 2012), which results in choices that are slower and more accurate than are optimal. Participants in the F-F condition were likely making decisions at a slower pace than would have been optimal, so sticky thresholds in C-F participants decreased performance because it moved them even further from the optimal speed. On the other hand, if C-C participants were already making slower-than-optimal decisions, sticky thresholds should have either (1) nudged F-C participants closer to optimal choice speed or (2) pushed them fully over the optimal choice speed into the "too fast" speed range. Results from phase 3 provide some indication that the latter possibility occurred: F-C participants earned less reward than C-C participants. These results speak to the strength of sticky thresholds.

One potential alternative explanation of these results is that performance level was incentive-specific. Participants in the F-F condition had the same amount of experience with the stimuli as C-F participants, but they had more experience *under the faster incentive*, so perhaps they outperformed C-F participants merely because this greater experience increased their accuracy relative to their speed, and not because they set a more appropriate

threshold. Results from the SAT model will address this issue and show that this alternate explanation has little support.

SAT Modeling Approach. Our second analysis examined speed and accuracy in a combined analysis, using the Linear Ballistic Accumulator (LBA) cognitive process model of decision making that accounts for the SAT (Brown & Heathcote, 2008). The LBA is a well-validated psychometric model of decision making that has been applied in a large range of contexts. SAT models like the LBA have been established as feasible models of performance in slow response time tasks, like the task studied here (e.g., Hawkins et al., 2014a, 2014b; Lerche & Voss, 2017).

The LBA model as applied to the task in Experiment 1 is schematically outlined in Figure 3, where panel A shows an example stimulus in the experiment. In this example, it is clear that option 2 is poor (high fixed price, high price per minute). However, it is less clear—at least from a quick inspection—which of options 1 and 3 is better, since option 1 has a lower fixed price but higher price per minute, and option 3 has a higher fixed price but lower price per minute. The upper row of Figure 3B shows the total cost of each phone pricing plan. We do not believe that participants reliably calculated the objective cost of each phone prior to making each choice. Rather we argue that participants obtained an overall impression for each phone, which we think of in terms of each phone's utility. Similar to traditional choice models, which divide utility into a deterministic and a random component, SAT models assume that every choice option has a deterministic utility (in this case, the overall price of each phone). This deterministic portion of utility determines the mean drift rate for each choice option, so evidence tends to accumulate more quickly in favor of higher-utility options than lower-utility options. The random portion of utility is introduced through noise in the decision process—trial-to-trial variability in drift rate. The deterministic portion of utility in this experiment took this form: if the total cost of the cell phone pricing plan for option i was x_i (Figure 3B – upper row), then we assumed the utility u of option i was $u(x_i) = exp(-x_i^{\beta})$, where β $[0, 1]$ is a free parameter representing sensitivity to the total cost of each phone pricing plan (Figure 3B – middle row;

for illustrative purposes, we assumed $\beta = .85$ in Figure 3B). The utility of each option was then normalized across the $j = 3$ phone options in each choice set to generate a drift rate for each option: $d_i = \frac{u(x_i)}{\sum_j u(x_j)}$ (Figure 3B – lower row); the drift rates therefore take a similar form to the choice probabilities of traditional choice models such as the multinomial logit. Drift rates for the set of j options on a given trial were sampled from a normal distribution with mean d_i and standard deviation s, which raced to threshold (b) from a starting point uniformly sampled between 0 and A, according to the LBA architecture (Figure 3C).

We performed quantitative model comparison between four models that each freely estimated from data a different set of parameters across stages of the experiment. The models differed in terms of which of the three key parameters—response threshold (b), sensitivity to pricing information (β), or non-decision time (τ)—was freely estimated in stages 1 and 2, and which parameters were constrained to a common value across stages of the experiment. Here, we use the term "stage 1" to refer to the two phases that took place before any change in context occurred for any condition (i.e., phases 1 and 2, above). We use "stage 2" to refer to the two phases that took place after a change in context occurred for two of the four conditions (i.e., phases 3 and 4, above).

The first model was a *null* model that assumed a single set of parameters was required to explain data in both stages of the task. This model contained 5 free parameters for each participant (b, β, τ, A, s), and primarily served as a baseline from which to assess performance of the three remaining models. The second model assumed that the *response threshold* parameter differed across stages but that the utility and non-decision time parameters did not (6 free parameters: $b_{\text{stage 1}}$, $b_{\text{stage 2}}$, β, τ, A, s). This model is interpreted as time pressure causing participants to be more or less cautious in their decisions across stages; if the hypothesized biased thresholds mechanism were underlying the observed data it would be reflected in support for this model. The third model assumed that the *sensitivity* parameter governing the shape of the utility function differed across stages but that the response threshold and non-decision time parameters did not (6 free parameters: b, $\beta_{\text{stage 1}}$, $\beta_{\text{stage 2}}$, τ, A, s). This model is interpreted as time pressure leading participants

A. Stimulus on an experimental trial

	Option		
	1	**2**	**3**
Fixed price:	\$42	\$60	\$49
Price per minute:	\$0.09	\$0.12	\$0.04
Usage:	700 minutes		

B. Model-based transformation of stimulus

	Option (*i*)		
	1	**2**	**3**
Total cost (x_i):	\$60	\$84	\$57
Sensitivity ($u(x_i)$):	7.95×10^{-15}	1.7×10^{-19}	3.17×10^{-14}
Drift rate (d_i):	**~.2**	**<.001**	**~.8**

C. SAT model race for response selection

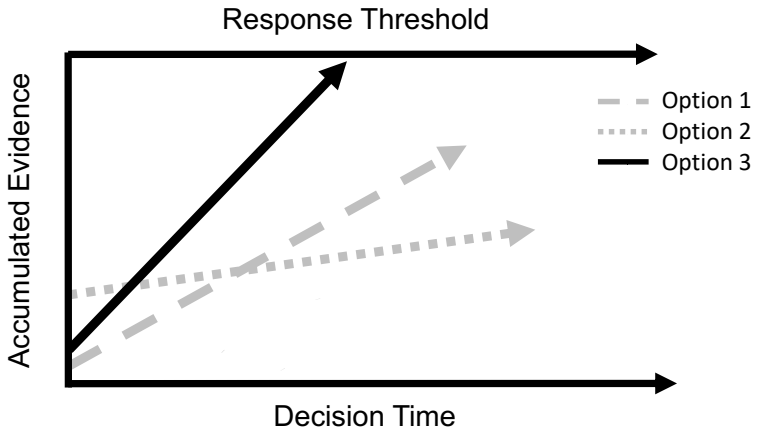

Figure 3. Schematic of the LBA model as applied to Experiment 1. See text for details.

to differentially interpret the cost differences among the phone pricing plans across stages. The last model assumed that the *non-decision time* parameter differed across stages but that the response threshold and utility parameters did not (6 free parameters: b, β, $\tau_{\text{stage 1}}$, $\tau_{\text{stage 2}}$, A, s). This model is interpreted as time pressure causing participants to change the speed with which they encoded the stimuli and made motor responses across stages, but the decision process was otherwise unaffected.

Our model selection procedure allowed us to determine which parameters provided the most parsimonious explanation of the data in each condition; the appropriate balance between model simplicity and goodness of fit to data, consistent with accepted best practices in cognitive modeling (Heathcote, Brown, & Wagenmakers, 2015). Specifically, we applied the 4 models independently to the 4 conditions of the experiment (F-F, F-C, C-F, C-C).[1] To quantitatively compare the set of 4 models applied to each condition, we used the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Lower DIC values indicate a better model for the data (i.e., the most parsimonious tradeoff between model complexity and goodness of fit to data). We conducted a parameter recovery study that demonstrated the parameters of each of the four models are identifiable in realistically sized data sets (Appendix B), which gives confidence in the model selection outcomes.

We used a hierarchical Bayesian framework to simultaneously estimate model parameters at the participant and group levels, which provides a principled way to incorporate individual differences and population level trends into the analysis. Model parameters were estimated using Particle Metropolis within Gibbs (PMwG; Cooper et al., 2021; Gunawan, Hawkins, Tran, Kohn, & Brown, 2020). All participant-level parameters were log transformed, except for the sensitivity parameter which was probit transformed, and estimated on the real line. Prior distributions for the group level mean parameters were distributed as

[1] We could also have investigated more complex models; for example, a model that allowed the response threshold and sensitivity parameters, but not non-decision time, to differ across stages, or to allow all three parameters to differ across stages. However, given the relatively few data points observed per participant in this experimental design, and initial explorations of these models with these data, such complex models could not be discriminated from the best model from the set of 4 models under investigation (i.e., model comparison indices were very similar, as was visual inspection of goodness of fit to data) and led to uninterpretable patterns in parameter estimates, so we did not pursue the complex models further.

multivariate normal with mean vector set to 0 for all elements and covariance matrix with diagonal elements set to 1 and off-diagonal elements set to 0. The group level variance-covariance matrix was set to the marginally non-informative prior distribution of Huang and Wand (2013). For further details on the PMwG prior distributions, see Cooper et al. (2021); Gunawan et al. (2020). Importantly, the prior distributions were equal across *stage* (1, 2) and condition (F-F, F-C, C-F, C-C) of the experiment, meaning that any differences observed in the posterior distribution of the parameters were driven by trends in data. As a guide for interpretation, sticky thresholds would be evidenced by a difference in population-level means in stage 2 such that response caution for the C-F condition is greater than response caution for the F-F condition, and response caution for the F-C condition is lower than response caution for the C-C condition.

Parameter estimation via PmWG involves three stages; we refer the reader to Gunawan et al. (2020) for the details and purpose of each stage and here we cover only the user-specified settings. In each stage we set the number of particles to 100. We first sampled 200 iterations as burnin so as to reach the target region of the parameter space. We then sampled up to 5000 iterations for developing an adaptive proposal distribution, which terminated early once enough unique samples were generated. In the final stage, we sampled and retained for analysis 5000 samples from the posterior distribution of the parameters. Convergence was confirmed by visual inspection of participant- and group-level chains.

Prior to estimating model parameters from data, we removed a small proportion of trials that were considered outliers. This procedure was performed to remove trials that were very unlikely to have been generated from the decision process under investigation—very fast anticipatory responses, very slow failing-to-attend responses—since these data points can bear strong and unwarranted influence on parameter estimation. We first removed any trials with responses that were slower than 30 seconds. Second, we removed trials that were inconsistent with the bulk of trials at the start or end of the distributions of response times (i.e., the fastest and slowest responses), independently for each participant. Specifically, we inspected the fastest 5 responses and if there was a difference of greater

than .8 seconds between two successive (very fast) responses, the faster of the responses was removed from analysis. Similarly, we inspected the slowest 2 responses and if there was a difference of greater than 5 seconds between two successive (very slow) responses, the slower of the responses was removed from analysis. This second procedure ensured that atypically fast or slow responses for a participant, given all of their other responses, were not included in the analysis. Together, the two exclusion criteria removed .49% of trials. We also removed from analysis participants who failed to complete at least 5 trials in a stage and participants who did not significantly exceed chance level performance (i.e., 33% accuracy) in the first or second stage of the experiment, which left 76 participants for the full model-based analysis. This step ensured that participants who simply pressed buttons very rapidly were not included in the SAT model analysis. Finally, to correct for partial guessing responses, and hence responses with very low likelihood under the LBA model, we assumed 5% contaminant responses in the predicted distributions of the model. For these contaminant responses, the response times were taken from a uniform distribution with a range defined separately for each participant, set to be equal to the range of the participant's response times observed in data, and response accuracy at chance level (33% Ratcliff & Tuerlinckx, 2002).

Data and code to estimate the SAT models as described here are available at `osf.io/wbyj7/`.

SAT Model Results. The DIC model comparison values are shown in Table 2. The difference between the lowest DIC (threshold model) and second-lowest DIC (non-decision time model), aggregated across conditions, was over 250 units; a DIC difference between models of 10 units is generally considered strong evidence in favor of the lower-DIC model (Pratte, Rouder, Morey, & Feng, 2010). Therefore, our results indicate very strong evidence for a model that allows the response threshold to differ across stages of the experiment, in the aggregate.

To ensure the DIC-preferred response threshold model sufficiently explains trends in the data and isn't simply the best-performing model of a poor set of models, we evaluate

Table 2
Experiment 1: DIC model comparison values. Bold entries in each row indicate the DIC-preferred model for the condition.

Condition	Null	Threshold	Sensitivity	Non-Decision Time
Fast – Careful	8405	**8030**	8387	8034
Careful – Careful	5741	**5709**	5679	5723
Careful – Fast	7634	**6975**	7516	7200
Fast – Fast	9223	**9028**	9156	9039
Sum	31003	**29742**	30738	29996

its descriptive adequacy. Figure 4 shows the group-level trends in observed and posterior predictive data across key summary statistics. The upper row summarizes the distribution of response times for correct responses with the 10^{th}, 50^{th} (i.e., median) and 90^{th} percentiles of the distribution; the middle row shows the median response time of the error responses, as there were too few data to obtain reliable estimates of the leading edge and tail of the distribution for visualization; and the lower row shows the mean accuracy.

The data (dots) fall within the 95% credible interval of the posterior predictive distribution for most summary statistics, indicating the SAT model provides a good description of the main trends in data. For example, the model provided a good account of the observed choice proportions in each of the two stages and four conditions (lower row – the symbols lie within the prediction interval of the model in most cases). Overall, the model also provided a reasonable account of the response time distributions. There are a few cases where the model predicted faster responses than people made (e.g., incorrect responses in the F-F condition), but critically this mostly occurred only for low probability responses (i.e., incorrect responses). The model also had some misfit for the conditions that had very large differences across stages. For example, from stage 1 to stage 2 the C-F condition decreased accuracy by 10% and almost halved the median response time. In this light, the SAT model's account of the data is impressive given that these behavioral changes were captured with just a single theoretical mechanism—a change in response threshold.

As expected, the model also picked up on key qualitative trends in data, such as the SAT. When instructed to switch from fast responding in stage 1 to careful responding in

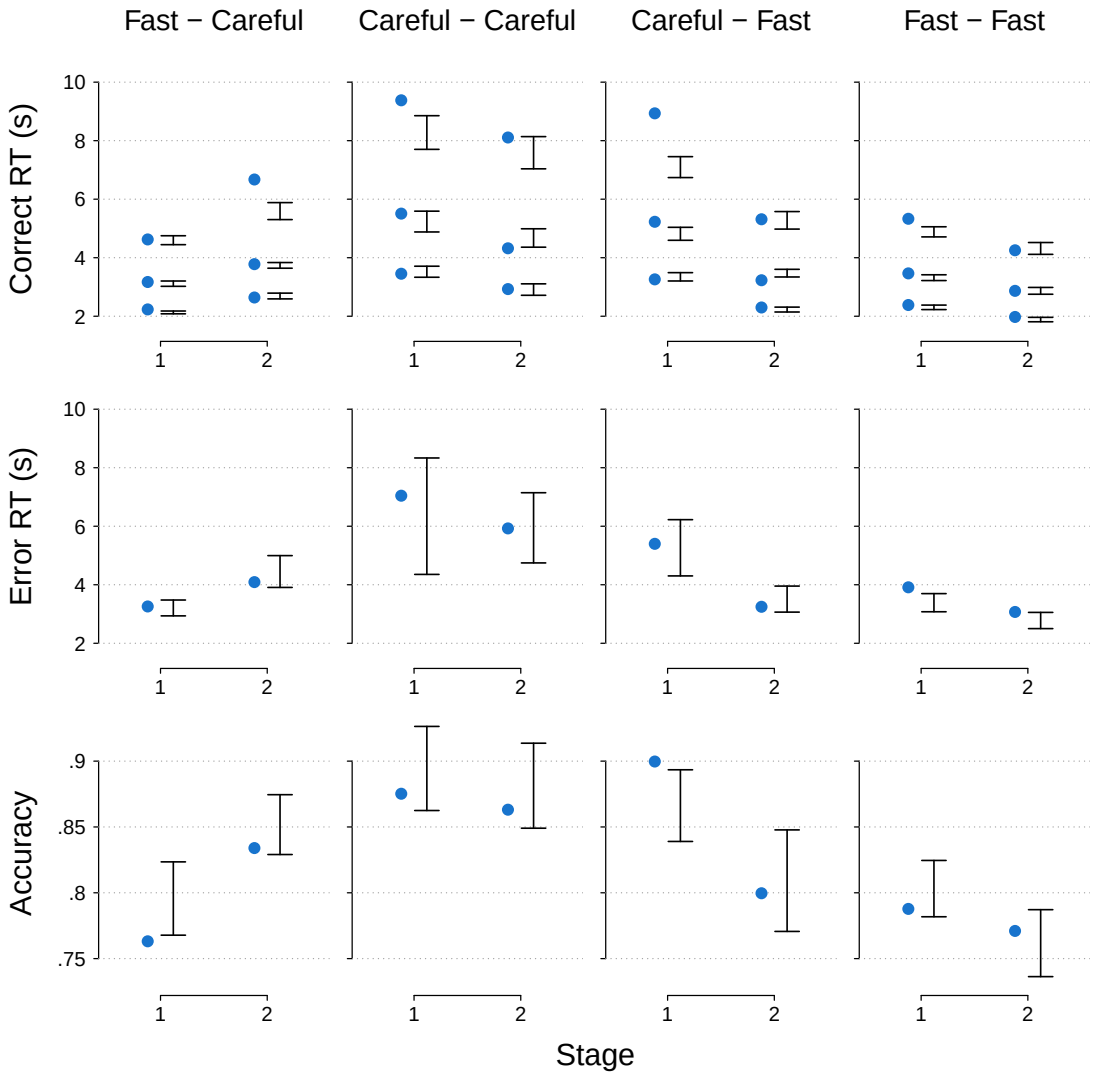

Figure 4. Experiment 1: Descriptive adequacy of the SAT model. The upper row shows response time (RT) distributions for correct choices. The vertically placed dots in each panel show the .1, .5 (median) and .9 quantiles of the RT distribution; symbols show data, uncertainty bars show the 95% credible interval of the posterior predictive distribution, and x-axis position shows the stage of the experiment. The middle row shows the response time distributions for incorrect choices where the dots and bars show the median. The lower row shows mean accuracy with the same display conventions as the response time panels.

stage 2 (first column), participants became more accurate but slower. In contrast, when instructed to switch from careful responding in stage 1 to fast responding in stage 2 (third column), participants became faster but less accurate.

The effect of the proposed biased thresholds can be seen in Figure 4. For example, when comparing the first two columns, the condition that received instructions that switched from fast (stage 1) to careful (stage 2) responding showed an increase in accuracy and slower response times in stage 2 (first column), as expected from the SAT, but the increase in accuracy and slowdown in response time was not as extreme as the condition that had been instructed to respond carefully in both stages. This indicates that responding in a fast regime had effects that persisted even when instructed to change to a more careful mode of responding; participants were unable to sufficiently disengage from the previous response regime.

To quantitatively explore this explanation of the data, we report the parameter estimates of the DIC-preferred threshold model in terms of the odds of a difference in the estimated parameter values between conditions, where larger odds indicate greater evidence for an effect (for similar approaches, see Hawkins, Hayes, & Heit, 2016; Mittner et al., 2014; Winkel et al., 2016). To calculate the odds for a given pairwise comparison of interest (e.g., response threshold between F-C vs. C-C in stage 2), we took the difference between the MCMC chains for the marginal posterior distributions of the respective group-level mean parameters, summed the number of posterior samples that were greater than zero, and converted this to the proportion of the total number of posterior samples that were greater than zero. The proportions p were converted to odds according to $p/(1-p)$, where p can be interpreted as the probability that a sample from one distribution is larger than a sample from another distribution.

Since odds indicate the strength of evidence for a given effect they have advantages over alternative metrics computed from posterior distributions such as credible intervals of difference distributions, which are based on all-or-none comparisons (i.e., a credible interval contains or does not contain the point of no difference). The odds therefore indicate the likelihood of a difference between two distributions relative to the likelihood that there was no difference. In this vein, odds are conceptually similar to the Bayes factor, which is commonly used to quantify evidence in Bayesian analyses, but the two statistics differ in

their assumptions about the role of prior distributions in determining the weight of evidence for an effect. Since we estimated parameters in a Bayesian framework, and we report odds as the evidence of the presence of an effect, we do not report conventional p-values. However, we can interpret odds as indicating positive evidence (>3:1), substantial evidence (>10:1), strong evidence (>30:1), or decisive evidence (>100:1) (cf. Jeffreys, 1967).

We present summary statistics and odds for the parameter estimates of the model in two parts, corresponding to the two primary comparisons of interest: participants who experienced the fast condition followed by the careful condition relative to those that experienced only the careful condition (i.e., F-C vs. C-C), and participants who experienced the careful condition followed by the fast condition relative to those that experienced only the fast condition (i.e., C-F vs. F-F). In particular, we were interested in whether the parameters of the model differed in stage 2 of the experiment where, for each of the pairs of interest (F-C vs. C-C, C-F vs. F-F), both conditions received the same task instructions. Therefore, any observed differences at stage 2 are attributable to carryover effects from the history of participants in the conditions. This will allow us to directly test for an effect of early experience on threshold levels.

Fast-Careful vs Careful-Careful. Table 3 provides the 95% highest density interval of the posterior distribution for the four model parameters, separated by condition: the response caution (reported separately for stages 1 and 2), the sensitivity parameter, and the non-decision time parameter. Here, we report the conventional measure of response *caution* in the LBA model rather than the raw response threshold (b). This is because the LBA model has two parameters related to the quantity of evidence required to trigger a response (A, b). The response caution measure deconfounds the two parameters through the transformation $b - A/2$, which is interpreted as the average amount of evidence required to trigger a response. Table 3 also provides the odds comparison for the difference between these parameters.

In stage 1 there was very strong evidence that participants instructed to respond fast were less cautious than participants instructed to respond carefully (>1000-to-1 odds).

Table 3
Experiment 1: 95% highest density intervals (HDI) and odds comparison for the group-level parameters of the SAT model.

		F-C	C-C	C-F	F-F
Caution – Stage 1	95% HDI	(1.33, 1.85)	(3.19, 4.55)	(2.33, 3.45)	(1.31, 2.09)
	Odds	>1000-to-1		>1000-to-1	
Caution – Stage 2	95% HDI	(1.75, 2.56)	(2.77, 3.98)	(1.27, 2.19)	(1.01, 1.66)
	Odds	>1000-to-1		11-to-1	
Sensitivity	95% HDI	(.81, .98)	(.82, .98)	(.80, .96)	(.76, .98)
	Odds	1.1-to-1		1.1-to-1	
Decision variability	95% HDI	(.24, .35)	(.21, .31)	(.24, .34)	(.23, .34)
	Odds	4.4-to-1		1.5-to-1	
Non-decision time	95% HDI	(1.04, 1.62)	(.63, 1.27)	(1.01, 1.55)	(.92, 1.52)
	Odds	27-to-1		1.7-to-1	

Interestingly, even though both groups were instructed to respond carefully in stage 2, participants who previously performed under fast instructions (F-C) were still less cautious on average than participants who had only performed under instructions to respond carefully (C-C; >1000-to-1 odds). This effect cannot be attributed to a failure of the F-C condition to adhere to task instructions and make more careful decisions from stage 1 to stage 2, as there was strong evidence that caution increased (226-to-1 odds). In contrast, there was no convincing evidence that the sensitivity or trial-to-trial variability in drift rates differed across the F-C and C-C conditions (1.1-to-1 and 4.4-to-1 odds, respectively), indicating that the shape of the utility function and decision variability did not differ between conditions. There was some evidence that F-C had a larger non-decision time than C-C (27-to-1 odds), indicating some difference in the speed of encoding the stimuli and producing a motor response between conditions.

Taken together, the pattern of parameter estimates confirms that participants performed the task in a manner consistent with the task instructions, where participants in the C-C condition made more cautious decisions than those in the F-C condition in stage 1. The difference in caution at stage 2, however, can only be due to differences in task history: F-C participants failed to appropriately increase their level of caution when tasked with switching from fast to careful responding. This provides evidence in favor of the hypothesized

biasing effect of the original threshold.

Careful-Fast vs Fast-Fast. Again, in stage 1 there was very clear evidence for a difference in response caution—participants who were instructed to respond carefully made more cautious decisions than participants instructed to respond fast, on average (>1000-to-1 odds). Similarly, when asked to respond fast in stage 2, there was substantial evidence for the hypothesized bias in threshold levels: participants who switched from careful-to-fast instructions did not decrease their caution to the same extent as participants who always responded under fast instructions (11-to-1 odds). As above, this effect cannot be attributed to a failure to adhere to the task instructions: there was very strong evidence that participants in the C-F condition made less cautious decisions in stage 2 compared to stage 1 (>1000-to-1 odds). There was no reliable evidence that the sensitivity, decision variability or non-decision time parameters differed across the C-F and F-F conditions (1.1-to-1, 1.5-to-1 and 1.7-to-1 odds, respectively).

Overall, the most striking difference from the previous set of comparisons is that the change in caution at stage 2 between the C-F and F-F conditions appears to be attenuated relative to the difference between the F-C and C-C conditions. This might be interpreted to mean that it is easier for people to switch from a regime of careful-then-fast responding than it is to switch from fast-then-careful responding.

Discussion. The conventional statistical analysis and SAT model analysis provided converging evidence in favor of the hypothesized biasing effect of early thresholds on later threshold levels. Participants who began the choice task with the incentive (and instruction) to make choices quickly set a relatively low threshold to enable fast decisions. When the incentive (and instruction) changed, participants raised their threshold in response, but the raising of the threshold was insufficient; we refer to this as sticky thresholds. As a result, these participants made faster (and less accurate) choices than participants who faced the same, careful incentives throughout all phases of the task. Sticky thresholds were also observed in participants who switched from a careful to a fast incentive. Direct evidence of biased thresholds came from the SAT model, which found threshold levels to be different, as

expected, between the focal comparison conditions. Indirect evidence from the conventional statistical methods corroborated these results. Participants switching from the fast to the careful incentives made faster and less accurate decisions than participants who always faced the careful incentive. Similarly, participants switching from the careful to the fast incentives made slower and more accurate decisions than participants who always faced the fast incentives.

The results of the SAT model provide evidence against the alternative explanation that the differences in performance were caused by a difference in skill at responding within a particular incentive condition. The model provided no convincing evidence that participants in the non-incentive-changing conditions were more skilled or efficient at extracting information from the stimuli than participants in the incentive-changing conditions (i.e., estimated values of the sensitivity parameter were not systematically different across the conditions). This is a strength of the SAT model as it allows us to uniquely attribute the observed pattern in behavior to a change in decision caution, and not to differences in sensitivity or non-decision time parameters. This kind of attribution is not possible in the conventional statistical analysis.

Experiment 1 provided evidence in favor of one proposed causal mechanism—shifts in perceptions of speed and accuracy. When participants faced a changed incentive that necessitated faster (and thus less accurate) choices, they appropriately made faster and less accurate choices. But what they perceived to be fast and inaccurate was not particularly fast or inaccurate, because their perceptions were benchmarked on earlier performance, which was slow and highly accurate. Evidence of these perception shifts came from participants' own reports of their speed and accuracy. Even in the presence of objective feedback about their speed and accuracy, perceptions of *fast* and *slow*, *accurate* and *inaccurate* were affected by early experience in the choice environment. This pattern of results is consistent with our theorizing, but it is only correlational, because we did not manipulate perceptions directly. Experiment 4 will provide more direct evidence of perception shifts. We next seek to replicate the biased thresholds effect while also providing evidence in favor of a second

causal mechanism: goal activation.

Experiment 2

We argue that after repeated experience in a particular decision environment, the environment itself activates prior goals. Therefore, when a decision maker attempts to set new goals in a decision environment, the environment will activate the prior goal, at least to some extent, thus preventing full adjustment and thereby leading to biased thresholds.

One particular property of goal priming that is critical in demonstrating its influence on thresholds in the SAT is that goals operate *across modalities* (Bargh & Chartrand, 1999). That is, if a particular goal is primed, such as the goal to achieve a high level of accuracy, then the goal is expected to be active for any choice faced by the decision maker; that is, the goal ought to be active both for choices made in the context and for choices made in other contexts. To test the hypothesis that goal activation influences thresholds, Experiment 2 studied a similar choice task as Experiment 1 and manipulated task instructions in a similar manner, then tested whether the goal of *fast* or *careful* responding transferred to an ostensibly unrelated perceptual discrimination task.

Method

The 169 participants in this study were members of a paid subject pool from a private university in the eastern United States. Participants engaged in a series of unrelated studies in a one-hour study session conducted in an on-campus computer lab in groups of 8 to 12. At the end of the one-hour session, participants received $10. A few days after the completion of the study, two participants received an email informing them that they had earned an additional $100 from the study.

The tasks studied in Experiments 1 and 2 were similar in structure though they differed in superficial features, which allowed us to test the generality of the observed effects beyond the specific task used in Experiment 1. Participants were shown the prices of four products from four different stores. Their task was to determine which of the four stores

provided the lowest overall price for the four products. The products were labeled, "Product 1", "Product 2", etc. The prices varied between $10 and $115, and were constructed such that the price range for any given product across the four options was at most $15, to reflect a realistic price range of a single product across several stores. As in Experiment 1, participants performed the task in four 2-minute phases. This experiment used two conditions, F-C and C-C, where the transition from Fast to Careful responding occurred in phase 3.

Instead of playing for money, participants were rewarded with points and were instructed that the two participants who earned the highest two scores would receive $100, which was determined after all 169 participants completed the experiment. In the Fast condition, participants were informed they would receive 10 points for every correct response and a 50-point bonus multiplied by their percentage correct. In the Careful condition, they were informed they would receive 2 points for every correct response and a 250-point bonus multiplied by their percentage correct. As in Experiment 1, participants made choices as quickly or as carefully as they wished within each two-minute phase. Participants received no feedback on their accuracy during the phases but were informed of their accuracy and points earned at the end of each phase.

To provide evidence of the role of goals in threshold anchoring, a seemingly unrelated task was added at the end of the fourth and final phase of the pricing task. This second task showed a 10 by 10 grid of white squares, many of which were filled with red dots. Participants were asked to judge whether the grid contained greater or fewer than 50 red dots (the grid never contained exactly 50 dots). They were instructed that they would be required to answer 25 of these trials. If F-C participants learned to associate the decision context with a stronger speed goal, then we would expect the experimental context to continue to activate the speed goal throughout the Careful phases of the pricing task and into the perceptual discrimination dot task. Participants received no additional monetary incentive for the grid task.

Table 4
Experiment 2: Mean response time and accuracy by condition and phase.

Average Response Time in seconds (standard error)

	Phase 1	Phase 2	Phase 3	Phase 4
F-C	4.9 (.09)	4.5 (.08)	6.6 (.14)	6.5 (.15)
C-C	9.1 (.23)	9.2 (.22)	9.5 (.24)	9.5 (.24)

Average Accuracy in percent (standard error)

F-C	55.8 (5.9)	54.8 (5.8)	63.1 (4.5)	68.3 (3.7)
C-C	70.9 (1.7)	76.1 (1.3)	77.0 (1.2)	77.0 (1.7)

Fast Condition

Results

Conventional Statistical Analysis. Table 4 provides the mean response time and accuracy, along with standard errors, for each condition and phase; the distribution of mean response times and accuracy rates are shown in Figure A1 (Appendix A). Data were analyzed using the same statistical methods as described in Experiment 1. (Also in keeping with Experiment 1, the reported means and standard errors are back-transformed to the original scale.) As predicted, participants in the F-C condition made faster responses than participants in the C-C condition in both phase 3 (M = 6.6 s vs. M = 9.5 s; $F(1, 668) = 151.7$, $p < .01$) and phase 4 (M = 6.5 s vs. M = 9.5 s; and $F(1, 668) = 151.8$, $p < .01$). This was also accompanied by significantly lower accuracy in both phases (Phase 3: M = 63.1% vs. 77.0%, $t(167) = 4.1$, $p < .01$; Phase 4: M = 68.3% vs. 77.0%, $t(167) = 2.7$, $p < .01$). This pattern of results provides additional evidence that the hypothesized bias in thresholds is not a fleeting phenomenon, as the bias continued through both post-incentive-change phases.

We propose that participants in the F-C condition made faster choices than C-C participants due, at least in part, to goal activation from the choice environment. If the choice environment activated a goal to make fast choices, then this goal should have remained active after choices in the original task were complete and subsequently transferred to the speed of any task immediately following. This goal activation implies that F-C participants

should respond more quickly to the perceptual discrimination dot task than C-C participants. Indeed, F-C participants responded significantly faster ($M_{F\text{-}C}$ = 1.8 s) than C-C participants ($M_{C\text{-}C}$ = 2.1 s; $F(1, 167) = 6.0$, $p = .02$) and were also less accurate (M = 72% vs. M = 75%; $z = 2.1$, $p = .02$).

We have posited three possible mechanisms underlying the biasing role of early decision thresholds on later threshold levels—perception shifts, goal activation, and decision strategy carryover. All three of these mechanisms may have had a role in the bias observed in this experiment, but only goal activation provides a satisfactory explanation of the difference in decision speed in the perceptual discrimination task, and thus it provides compelling evidence in favor of the role of goal activation in biasing threshold levels. Perception shifts are an unlikely explanation of speed differences in the dot task because the average response time in that task was over three times faster than the fastest average response time from the store price task. This makes it unlikely that perception shifts in the evaluation of speed caused the difference in response times. In addition, the psychophysics literature has repeatedly found that perceptions are context-specific (Stevens, 2017). The context-specific nature of these evaluations implies that participants' evaluations of speed and accuracy in the store price context are unlikely to have affected their perceptions of speed and accuracy in the perceptual discrimination task. In contrast, goals are known to affect behavior across contexts (Bargh & Chartrand, 1999). Finally, decision strategy carryover cannot explain the speed differences in the perceptual discrimination task because no conceivable decision strategy would apply to the store price task and to the dot task, as they share no similarity in structure.

SAT Model Approach. The SAT model for the pricing task was implemented in the same manner as described in Experiment 1 with the following exceptions. There were 4 stores, so each choice was composed of a race between $n = 4$ accumulators, and each store contained 4 products, so the drift rate for each accumulator was given by the sensitivity to the summed pricing information of the 4 products available in each store. As above, for the SAT model we refer to stage 1 (corresponding to phases 1 and 2 of the experiment) and

stage 2 (corresponding to phases 3 and 4 of the experiment).

We also applied the SAT model to the perceptual discrimination task to directly test the hypothesis that the store pricing task activates goals in the subsequently experienced perceptual judgment task. We did this with a so-called 'joint model' in that the two tasks were modeled in parallel in the same framework, which allows us to examine parameter associations across tasks (e.g., Wall et al., 2021). For the perceptual discrimination task, we assumed that the psychophysical representation of the perceptual stimulus was mapped to evidence accumulation rates in the SAT model via a cumulative normal link function (for similar approach, see Vandekerckhove, Tuerlinckx, & Lee, 2008). Given a stimulus with n dots, the drift rate for the response 'greater than 50 dots' was $d_{>50} = \Phi(\frac{n-\mu}{\sigma})$, where $\Phi(.)$ is the CDF of the standard normal distribution, and μ and σ are free parameters representing the mean and standard deviation of the cumulative normal link function. The drift rate for the response 'fewer than 50 dots' was $d_{<50} = 1 - d_{>50}$. The mean of the link function represents the location at which the participants switch from an average response of fewer than 50 dots to greater than 50 dots, which, due to perceptual biases, may not accurately reflect the true midpoint of the stimulus range (50 dots). The standard deviation of the link function represents the scale of the perceptual representation. Larger values indicate greater variability in the psychophysical representation of number, which indicates poorer ability to discriminate greater from fewer than 50 dots.

To this end, we freely estimated the location (mean) and variability (standard deviation) of the link function. Our use of the link function simplified the broad stimulus range (30-70 dots) via a plausible psychophysical mapping to a response function in the SAT model. The remaining SAT model parameters for the perceptual discrimination task were freely estimated from data; response threshold, starting point of evidence accumulation, across-trial variability in drift rate, and non-decision time.

All details including the 4 models to be compared in the pricing task, model specification, prior distributions, and parameter estimation were conducted using identical methods as described in Experiment 1. In the perceptual discrimination task, the group-level prior

distribution for the mean was centred at 0 for all parameters except for the location (mean) of the psychophysical function, which was set to 3.91 (i.e., log(50) – the midpoint of the perceptual range). All other prior settings were as described in the pricing task.

SAT Model Results. The DIC model comparison values are shown in Table 5. The difference between the lowest DIC (threshold model) and second-lowest DIC (non-decision time model) was approximately 480 units, again indicating very strong evidence that the model allowing response thresholds to differ across stages of the experiment provided the best explanation of the data.

Table 5
Experiment 2: DIC model comparison values. Bold entries in each row indicate the DIC-preferred model for the condition.

Condition	Null	Threshold	Sensitivity	Non-Decision Time
Fast – Careful	32215	**30849**	32210	31163
Careful – Careful	21350	**21127**	21330	21291
Sum	53565	**51976**	53540	52454

Figure 5 shows the SAT model's descriptive adequacy for the pricing task in the same format as Figure 4. The model provided a good explanation of all qualitative and quantitative trends in the pricing task: the switch to more accurate but slower responses in stage 2 relative to stage 1 in the F-C condition (left panel). Importantly, the change in the SAT in the F-C condition did not appear to approach the level of caution shown by participants in the C-C condition (right panel), who were still more accurate and slower in stage 2, even though both conditions received the same task instructions. The lower two rows in Figure 5 show that the SAT model also explained the trends in the perceptual discrimination task. Responses were fastest at the extremes of the perceptual range—very few or very many dots in the display. These responses were also very likely to be correct, shown in Figure 5 as always responding <50 in trials with very few dots and >50 in trials with many dots. Decisions in the middle of the stimulus range are considerably more difficult, indicated by slower responses and response proportions much closer to 50:50. Although not directly displayed in Figure 5, when collapsed across number of dots the posterior predicted accuracy and mean response time were both greater for the C-C condition than the F-C condition,

as observed in data.

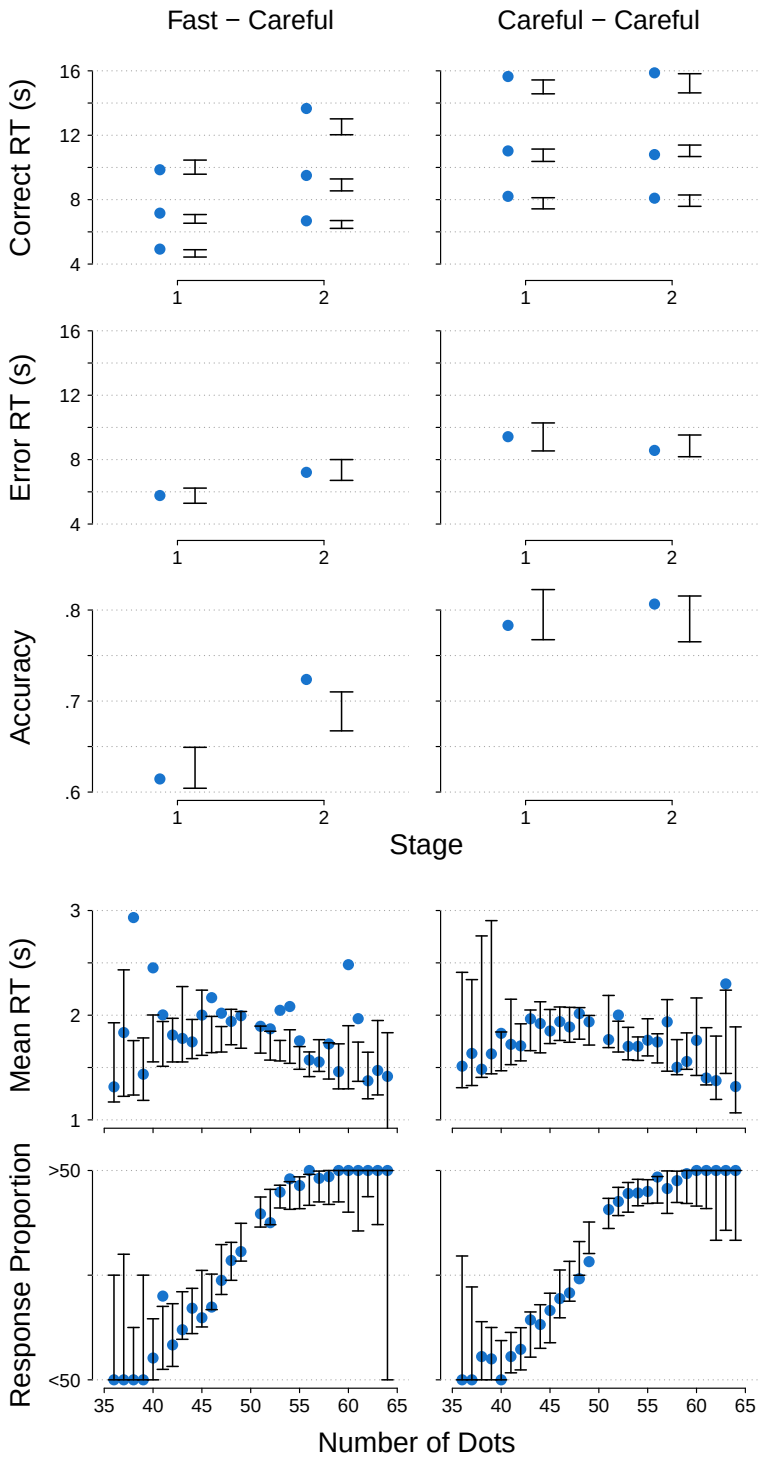

Figure 5. Experiment 2: Descriptive adequacy of the SAT model for the pricing task (upper three rows) and the perceptual discrimination task (lower two rows). The pricing task observed and posterior predictive data are displayed in the same format as Figure 4. For the perceptual discrimination task, mean RT (fourth row) and response proportion (fifth row) are shown as a function of the number of dots in the stimulus display. Within each panel, the dots show data and uncertainty bars show the 95% credible interval of the posterior predictive distribution.

Table 6 shows the 95% highest density interval of the posterior distribution of the group-level mean parameters of the SAT model, separated by task (rows) and experimental conditions (columns). In stage 1 of the pricing task, when task instructions differed between conditions, there was very clear evidence that participants in the C-C condition had greater response caution than participants in the F-C condition (>1000-to-1 odds). In stage 2, when task instructions were identical between conditions, there was strong evidence that participants who previously performed under fast instructions (F-C) made less cautious decisions than participants who only performed under instructions to respond carefully (C-C; 356-to-1 odds). As in Experiment 1, this effect could not be attributed to a failure to attend to task instructions; there was strong evidence that F-C participants had greater caution in stage 2 compared to stage 1 (>1000-to-1 odds). In contrast to Experiment 1, there was evidence that the sensitivity and non-decision time parameters were larger for C-C than F-C participants (both >1000-to-1 odds). This likely reflects that, across both stages, C-C participants were more sensitive to differences in store prices, and were slower to produce a motor response to indicate their choices. This collection of parameter differences is a hallmark of decision making that emphasizes speeded performance over accurate performance, which indicates that goal activation in this particular experiment affected more than just the thresholds (e.g., Dutilh et al., 2019; Rae et al., 2014).

In contrast to expectations, in the perceptual discrimination task there was no evidence for different levels of response caution between conditions. Instead, our strongest finding was that the F-C condition had a noisier representation of the perceptual display (variability in the psychophysical function; 99-to-1 odds). This suggests that the F-C condition had weaker encoding of stimulus information than the C-C condition, which is consistent with the lower sensitivity to stimulus information found in the pricing task. We also observed faster non-decision times for the F-C condition (49-to-1 odds). The patterns in these two parameter results explain the qualitative effects observed in data: faster non-decision time for the F-C condition will lead to faster mean response times compared to the C-C condition, and noisier perceptual representation will lead to lower accuracy. Finally,

Table 6
Experiment 2: 95% highest density intervals (HDI) and odds comparison for the group-level parameters of the SAT model.

			F-C	C-C
Store Pricing	Caution – Stage 1	95% HDI	(1.79, 2.63)	(3.65, 4.53)
		Odds	>1000-to-1	
	Caution – Stage 2	95% HDI	(2.42, 3.57)	(3.64, 4.63)
		Odds	356-to-1	
	Sensitivity	95% HDI	(.61, .66)	(.68, .71)
		Odds	>1000-to-1	
	Decision variability	95% HDI	(.11, .15)	(.11, .15)
		Odds	1.2-to-1	
	Non-decision time	95% HDI	(1.66, 2.43)	(3.32, 4.44)
		Odds	>1000-to-1	
Perceptual Discrimination	Caution	95% HDI	(.59, .73)	(.58, .76)
		Odds	1.7-to-1	
	Psychophysical function – location	95% HDI	(45.42, 47.26)	(46.29, 47.77)
		Odds	6.6-to-1	
	Psychophysical function – variability	95% HDI	(5.84, 8.37)	(3.40, 6.06)
		Odds	99-to-1	
	Decision variability	95% HDI	(.26, .35)	(.30, .41)
		Odds	21-to-1	
	Non-decision time	95% HDI	(.59, .72)	(.68, .81)
		Odds	49-to-1	

we also observed an effect of decision variability across trials such that the F-C condition made less variable decisions than the C-C condition (21-to-1 odds).

Discussion. Experiment 2 replicated the biased thresholds effect found in Experiment 1 for the pricing task: participants in the F-C condition made faster and less accurate choices than participants in the C-C condition, and our SAT model analysis demonstrated that this effect arose because F-C participants made less cautious choices (i.e., had lower response thresholds) than C-C participants. Participants in the F-C condition appeared to set a goal that emphasized speed of choices. This speed goal continued to affect decisions throughout the latter phases of the same task, when the incentive structure had changed so as to reward slower decisions, and even transferred to an ostensibly unrelated perceptual discrimination task. However, the stronger speed goal adopted by F-C participants manifested in an unexpected way. Rather than decreasing their response threshold

(i.e., decreasing their response caution), the stronger speed goal decreased participants' non-decision time and increased variability of their psychophysical representation (i.e., a decreased ability to accurately assess the number of dots). Overall, the results provide evidence in favor of the role of goals in the speed-accuracy tradeoff, but the unexpected manifestation of speed goals in non-decision time and in the variability of the psychophysical representation rather than in response thresholds weakens the evidence of the influence of goals on thresholds. To strengthen this evidence, Experiment 3 provides direct evidence of our proposed mechanism—that the decision context itself activates the decision goals that were active during prior decision occasions.

Experiment 3

We have proposed goal activation as one mechanism leading to biased decision thresholds in changing choice environments. After multiple decisions in a decision context, the learned decision goals become automatically activated by the decision context (Bargh & Barndollar, 1996). When a change in a decision context requires an adjustment to the threshold, we hypothesize that the decision context continues to activate the previous balance of speed and accuracy goals, thereby influencing decision thresholds. Experiment 2 provided evidence of the influence of decision goals via differential decision speeds in an unrelated choice task. Experiment 3 seeks to provide more direct evidence of goal activation by documenting its influence within the same decision context separated over different days.

Method

Forty-seven participants from a Western University in the United States took part in Experiment 3 in return for extra credit in an undergraduate business class and a monetary reward based on their performance on the experimental task. The experiment was conducted online in two parts, with a one-day separation between the two parts. Experiment 3 utilized the same stimuli as Experiment 1, a cell phone pricing task (two-tier prices with a base price

for usage under 500 minutes and a per-minute charge for minutes beyond 500). In part 1, participants engaged in two 2-minute phases of the task. In the Fast condition, participants were paid \$0.50 times their accuracy plus \$0.07 for each correct response. In the Careful condition, participants were paid \$1.50 times their accuracy plus \$0.01 for each correct response. To help participants determine the optimal effort level, they were provided with a table that showed the resulting earnings from various levels of speed and accuracy.

At the conclusion of part 1, participants entered their email address. Twenty-four hours later, participants received an email with a link to part 2 of the experiment, in which participants engaged in two more 2-minute phases of the task. In these last two phases, all participants were informed that they would be paid \$1.00 times their accuracy plus \$0.04 for each correct response. We hypothesized that those in the Fast condition in the first two phases would respond more quickly in the final two phases than those in the Careful condition. This would establish the hypothesized mechanism, as the original goals could not last an entire day. Instead, the decision context itself must have newly activated previous speed or accuracy goals.

To control for perception differences between the conditions, the difficulty of the first two phases was manipulated to ensure that participants in both conditions made guesses at the same average speed. In the Fast condition, 30% of the cell phone pricing plan choices were "easy", while 70% were "difficult" (to induce slower choices despite the faster incentive). In the Careful condition, these percentages were reversed (to induce faster choices despite the slower incentive). An "easy" choice is defined as one where usage was under 500 minutes (thus requiring only examination of the base prices) or the existence of a dominating alternative (with both the lowest base and variable rate prices). A "difficult" choice is one with usage over 500 minutes without a dominating alternative. We hypothesized that the difference in difficulty level would lead to identical average decision speeds despite the difference in goals. Nevertheless, balancing the average decision speed in this way alters the ability of the SAT model to unambiguously test our hypothesis of sticky thresholds between the Fast and Careful conditions because two experimental factors vary between

the conditions; speed incentives changed, which was our primary focus, but so too did the distribution of decision difficulty, which was required to balance average speed. Therefore, we cannot attribute potential parameter effects to a specific cause. For this reason, we do not analyze the data of Experiment 3 with the SAT model, a point we return to in the Discussion. In the two test phases in part 2 of the experiment, which were common to all participants, 50% of choices were easy. No accuracy feedback was given until the end of all four phases, whereupon participants answered a few additional questions about the task.

At the conclusion of part 2, participants were informed of their earnings and received instruction on where they could collect those earnings. Of the 47 participants who completed part 1 of the study, six did not complete part 2 after multiple email reminders, which left data from 41 participants.

Results

As hypothesized, mean decision speed was not significantly different between the Fast (M = 6.56 s) and Careful conditions (M = 6.23 s) during part 1 of the experiment, t(39) = -.35, p = .64. Thus, differing speeds in part 2 of the study were not caused by differences in perceptions of speed, because both conditions engaged in the same average speed. As no accuracy feedback was given, we can also rule out learned accuracy evaluations. To further ensure that the change in the proportion of easy decisions was not noticed by participants, all participants answered the following question at the end of their experience: "Making guesses seemed to get easier in the last two phases than the first two phases." For participants in the Fast condition guesses did get easier, but their agreement with this statement (M = 4.3) was not higher than that of Careful condition participants (M = 4.4), t(39) = .53, p = .60.

As differences in speed and accuracy perceptions are ruled out as an explanatory mechanism, any differences in speed in part 2 of the study are likely due to differential goal activation by the decision context. In phase 3, at least one day later, participants in the Careful condition made decisions more slowly (M = 6.52 s) than participants in the Fast

condition (M = 5.81 s), t(39) = 2.63, $p < .01$. This difference remained in phase 4 (M = 6.14 s vs. M = 5.40 s), t(39) = 2.93, $p < .01$. The decision goals that were activated in the first two phases of the task were reactivated by the decision context over a day later, causing differences in decision speed between the conditions. This difference occurred even though both conditions were explicitly instructed with the same incentive structure in final two phases.

Discussion

We propose that in repeated decisions, the decision context itself activates a balance of speed and accuracy goals that determines the decision threshold set by the decision maker. Experiment 3 provides evidence of goal activation by the decision context. Because of the one-day delay between part 1 and part 2 of the experiment, the difference in speed between the two conditions must have been produced by goals newly activated by the decision context. The average speed of participants in part 1 was equalized across conditions, so differing perceptions of speed could not have produced the resulting speed differences. The results of Experiment 3 also serve to generalize the effects proposed in this paper to the time-frame observed in real-world decisions. For pragmatic purposes, the decisions observed in Experiments 1 and 2 were made in quick succession. However, Experiment 3 separated the decision phases by a day or more, and the effects remained. The learning associated with a decision context is preserved from one decision occasion to the next, even when those occasions are separated by hours or days.

One weakness of Experiment 3 was that the structural differences between the conditions (i.e., differing choice difficulty levels) in phases 1 and 2 resulted in different learning, and thus accuracy did not follow the same pattern as in Experiments 1 and 2. Similarly, the differential learning in phases 1 and 2 means that it was not sensible to analyze the data from Experiment 3 with the SAT model as it was specified in Experiments 1 and 2—structural differences in task context—so it was not explored. Appendix C provides additional analyses of Experiment 3 and a discussion of those results, which reinforces our

decision not to analyze these data with the SAT model.

Perceptions of decision speed are proposed to influence decision thresholds. Experiment 1 provided evidence for this mechanism through participants' reported perceptions of decision speed (and accuracy). However, because perceptions were not manipulated directly, alternative explanations can be proposed for the results. Experiment 4 manipulates perceptions of decision speed directly, thereby providing more direct evidence of the role of perceptions in influencing decision thresholds.

Experiment 4

Experiment 4 had two aims. First, we sought to provide direct evidence of the role of perceptions in threshold levels. A second goal was to provide additional evidence that decision strategy carryover cannot by itself account for the effects obtained.

Method

Ninety-five participants from a private university in the western United States took part in Experiment 4 in return for extra credit in an undergraduate business class. During the study, participants learned that they would also receive two candy bars and would be entered into a drawing for $25. The study was conducted in an on-campus computer lab in sessions of 15 to 20 students. After checking in with a research assistant, participants were shown to a computer. Once all participants in a session had been checked in, the research assistant showed participants a URL to begin the study. The entire study was conducted from this URL without further instructions from the research assistant. At the end of the study, participants were checked out by a research assistant, who also gave them two candy bars chosen during the study.

Participants used the Mouselab procedure (Payne et al., 1988) to make 15 candy bar choices, two of which were randomly selected and given to the participant at the conclusion of the experiment. In the Mouselab procedure, attribute descriptions of the choice options are hidden until the participant moves the mouse cursor over that attribute. The timing

and pattern of mouse movements are recorded to enable the researcher to examine the participant's decision process and decision speed with a variety of measures. Each choice displayed three candy bars described on four attributes: name (e.g., "Snickers"), flavor (e.g., "Chocolate, nougat, peanuts, caramel"), size (e.g., "Regular"), and ounces (e.g., "2.07 oz."). The four attributes were displayed vertically (in rows) and the three candy bars were shown horizontally (in columns). Figure 6 shows an example trial.

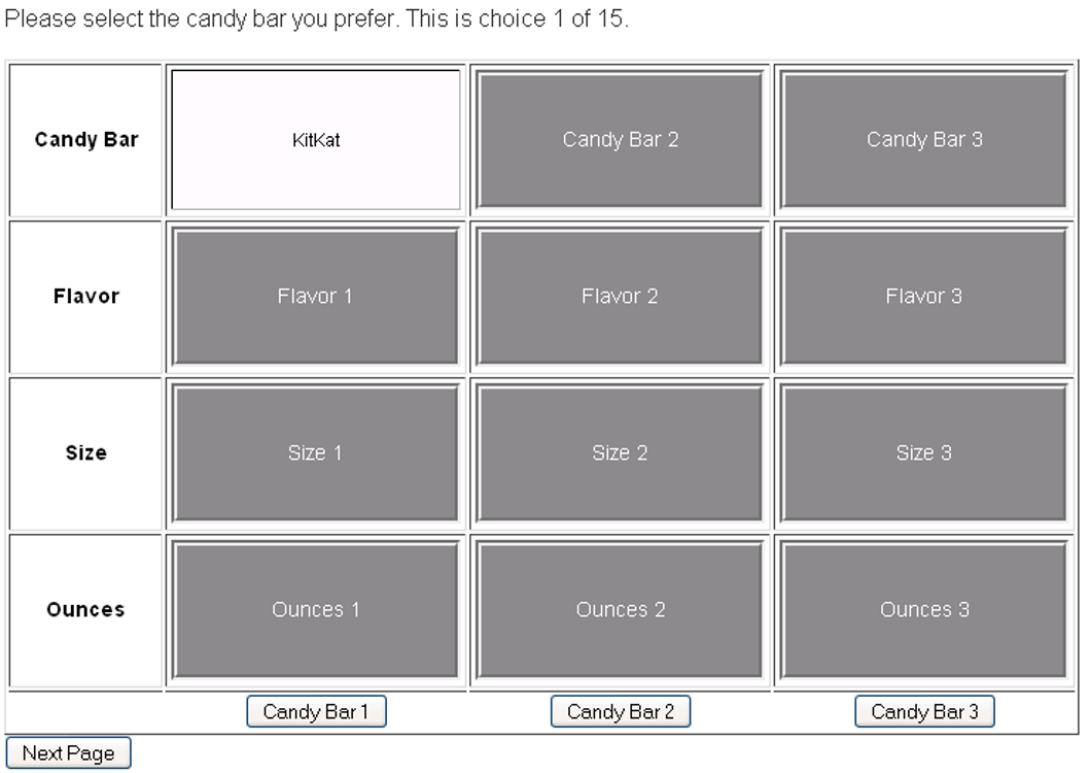

Figure 6. Experiment 4: Example trial. The participant's mouse is currently placed over the "candy bar" attribute of option 1, displaying the value for that attribute. All other attribute values are occluded until the mouse is placed over them.

Prior to commencing the task, participants were informed that they would be choosing candy bars and would receive two candy bars from those choices. The Mouselab choice procedure was then described to them, and they were told that they would have the opportunity to practice it. The practice phase required participants to examine three columns of numbers and determine which column had the highest sum. The columns of numbers were

arranged in the same size and orientation as the candy bar choices would later be arranged. Participants were informed that those who correctly identified the column with the highest sum in all five choices would be entered into a drawing for $25.

Speed perceptions were manipulated in these five practice decisions. In the Careful condition, each box contained a number randomly generated between 1 and 9. In the Fast condition, the randomly-generated numbers in each column were restricted to between 1 and 3, 4 and 6, or 7 and 9 (each group of numbers randomly assigned to one column). Thus, in the Fast condition, a quick glance at any of the numbers in the three columns was sufficient to inform participants of which column contained the highest sum. In contrast, in the Careful condition, participants were required to carefully examine each list to make the correct response.

We hypothesized that participants in the Careful condition would spend more time making their candy bar choices. Because they had expended more time to identify the highest sum, participants in the Careful condition would have different perceptions of speed than would participants in the Fast condition. As a result, participants in the Careful condition would make slower candy bar choices, despite having an identical incentive to progress quickly.

An alternative explanation for this result could be a carryover of decision strategies. That is, one might argue that in the Careful numbers condition, participants learned a slower decision strategy, and subsequently applied this slower strategy toward making the candy bar choices. The Mouselab procedure provides the additional benefit of allowing examination of decision strategies, to ensure that a perception shift in the evaluation of speed, and not decision strategy carryover, is the underlying cause of the slower decisions.

Results

The measures used to summarize a participant's decision process from a Mouselab choice can be classified into two categories: those that measure speed and those that measure processing style (for a review of measures, see Payne et al., 1988). The three measures of

decision speed are 1) BoxTime (the total amount of time spent with the mouse cursor over a box); 2) TimePerAcquisition (the average time spent on each box); and 3) Acquisitions (the total number of times the mouse cursor was moved over a box). The natural log of each measure was taken to eliminate right skew.

As a manipulation check, we examined decision speed during the numbers practice task. As expected, when it was more difficult to determine the column with the highest sum, participants were slower on all measures: BoxTime (M = 24.1 s vs. M = 11.9 s; t(93) = -12.99, $p < .001$), TimePerAcquisition (M = .76 vs. M = .49; t(93) = -10.63, $p < .001$), and Acquisitions (M = 31.8 vs. M = 24.5; t(93) = -6.52, $p < .001$).

The difference in speed observed between conditions in the numbers task caused a difference in speed perceptions, which caused a difference in the speed of candy bar choices. Participants in the Careful condition were slower than participants in the Fast condition on two of the three speed measures: BoxTime (M = 10.2 vs. M = 9.4; t(93) = -2.71, $p = .008$) and Acquisitions (M = 25.0 vs. 23.5; t(93) = -2.55, $p = .012$), but TimePerAcquisition did not significantly differ (M = .41 vs. M = .40; t(93) = -1.12, $p = .266$); a MANOVA test of the three variables was significant (F(2, 279) = 3.7, $p = .026$).

We proposed that the observed differences in speed across conditions were due to differences in the perception of speed; that is, what people think is a *fast* or *slow* response. A potential alternative explanation is carryover in decision strategy: participants in the more challenging (Careful) numbers choice task might have adopted a more careful decision strategy than participants in the easier (Fast) numbers task, and then applied this same careful decision strategy to the candy bar choice task. We used three measures of decision strategy derived from the Mouselab procedure: 1) VarAtt (the variance in the proportion of time spent examining each attribute); 2) VarAlt (the variance in proportion of time spent examining each alternative); and 3) Pattern (the proportion of attribute-based versus alternative-based transitions). In the numbers task, the difference in speed between the conditions was not accompanied by any significant difference in VarAtt (M = .15 vs. M = .16; t(93) = -1.63, $p = .11$) or VarAlt (M = .20 vs. M = .18; t(93) = 1.58, $p = .12$).

A significant difference was found in Pattern (M = .59 vs. M = .68; t(93) = -5.1, $p <$ .01), which indicates some difference in decision strategies between the two conditions in the numbers practice task. Critically, however, the difference in Pattern did not carry over to the candy bar choice task, and in fact it reversed (M = .19 vs. M = .14; t(93) = 3.63, $p <$ .01). Again, no significant differences were found in VarAtt (M = .31 vs. M = .30; t(93) = .75, p = .45) or VarAlt (M = .20 vs. M = .20; t(93) = -.21, p = .83). Bayesian t tests on VarAtt and VarAlt gave Bayes factors of 3.61 and 4.55 in favor of the null hypothesis, thereby providing positive evidence that decision strategies did not differ across conditions. While these process measures are not perfect reflections of underlying decision strategies, the results provide some evidence that decision strategy carryover is not a contributing cause of the observed differences in speed.

We have proposed that one cause of sticky thresholds is a shift in perceptions of speed and/or accuracy. Experiment 4 manipulated participants' perceptions of choice speed by giving them easy or difficult choices during training. Faster (vs. slower) choices in training created a benchmark for speed that resulted in faster speeds being perceived as normal. This difference in speed perceptions led to a difference in choice speeds in subsequent choices for candy bars. The Mouselab procedure enabled granular examination of processing patterns, which gave no indication that decision strategies contributed to the difference in choice speeds, thus attesting to the role of perceptions of speed in decision thresholds.

A Model of Threshold Adjustment

The SAT models examined in Experiments 1 and 2 utilized a single parameter for a participant's threshold under a particular incentive structure. In other words, the model assumes each participant employs the same threshold on all decisions within the same reward scheme. In reality, decision makers may adjust their threshold after each decision based on their perception of the length of time the decision required, whether their decision was correct (or their confidence in the correctness of the decision), or both. The experiments we have presented were designed to investigate the macro-level phenomenon of threshold

underadjustment, not the decision-to-decision adjustment process, and thus the data are unlikely to enable investigation of this micro-level process. However, here we propose and simulate one such micro-level process to illustrate a plausible mechanism that may give rise to threshold underadjustment.

We simulate four decision makers—one in each experimental condition—who each make 2000 total decisions. The simulation reported here explored many more trials than participants completed in our experiments. This is to demonstrate that our proposed adaptation mechanism captures the enduring effect of previous incentive structures on current choices. Nevertheless, when the mechanism is simulated with fewer trials to more closely mimic the experiment structure (∼100) it still produces all qualitative effects observed in the experiment, though with the expected increase in variability across simulation runs owing to the smaller number of trials.

Each decision was generated by the same LBA process assumed in the SAT models studied earlier. Each decision maker is assigned to one of the conditions of Experiment 1: Fast-Fast, Fast-Careful, Careful-Fast, and Careful-Careful. The decision makers adjust their thresholds after each decision based on three criteria: (1) the incentive structure, (2) the perceived speed of the decision, and (3) whether the previous decision was correct or incorrect. After correct decisions, the simulated participants tend to lower their threshold to enable faster decisions, and after incorrect decisions the participants increase their threshold to increase the likelihood of correct decisions in the future. The size of these adjustments depends on the incentive structure and the perceived speed of the decision, as depicted in Table 7.

Table 7
Simulation: Threshold adjustments as determined by incentive, perceived speed, and correctness.

Fast Incentive

		Result	
		Correct	Incorrect
Perceived Speed	Fast	-.1	.1
	Slow	-.3	0

Careful Incentive

		Result	
		Correct	Incorrect
Perceived Speed	Fast	0	.3
	Slow	-.1	.1

Under the fast incentive structure participants aim to make fast decisions at the expense of accuracy. To facilitate this decision making pattern, we argue that participants are more likely make larger downward shifts in their thresholds than upward shifts. When the previous decision was correct, participants always lower their thresholds for the next decision because they may be able to respond faster and still be correct. However, correct decisions that were already perceived to be fast lead to only a small decrease in the threshold, while a correct decision that was perceived to be slow produces a larger threshold decrease. In contrast, threshold adjustments following incorrect decisions depend on the perceived speed of the decision. Incorrect but fast decisions cause the participant to slightly increase their threshold, because the error may have been due to an overly-hasty response style. However, participants in the fast incentive structure were assumed to make no threshold adjustments following errors that were perceived to be slow, because this goes against the fast incentive motivation.

Under the careful incentive structure participants aim to make careful decisions with lesser regard for the decision speed. To facilitate careful responding, we assume participants made the same magnitude of threshold adjustments as the fast incentive structure but in the opposite direction. Thresholds are always increased following an incorrect decision because errors are inconsistent with the careful incentive. Errors that were perceived to be fast lead to a large threshold increase, while errors that were already slow increase the threshold only marginally; in the careful incentive structure a fast error is a more egregious violation of the instructions, hence the larger adaptation. Correct decisions that were perceived to be slow lead to a small decrease in threshold, because it may have been feasible to respond more rapidly and still be correct. However, correct yet fast decisions don't lead threshold adjustments due to the risk of responding too hastily.

Participants' perceptions of decision speed evolve throughout their experience in the decision task following a Bayesian updating process. We assume participants' perceptions of decision speed are represented with a Gamma distribution, where $\Gamma(a, b)$ represents the Gamma distribution with shape a and rate b. We assume that participants begin the task

with a $\Gamma(5, 1)$ prior distribution, which has a 5 second mean decision time. Decisions faster than 5 seconds (i.e., faster than average) are perceived to be fast and decisions slower than 5 seconds are perceived to be slow. The decision speed sampling distribution is updated after each decision following $\Gamma(5, \beta)$, where β is the update parameter. As participants in the Fast incentive make faster decisions, their perceived mean decision speed—the delineation between a fast and slow decision—will tend to decrease. Similarly, as participants in the Careful incentive make slower decisions, their perceived mean decision speed will tend to increase.

At the halfway point of the experiment (after 1000 decisions in our simulation), participants in the Fast-Careful and Careful-Fast conditions change incentive. Their perceptions of their decision speed continue to evolve according the same Bayesian updating mechanism.

Figure 7 shows the time series for the evolution of thresholds according to our threshold adjustment mechanism for the exemplar participants in each condition of Experiment 1. In the Fast-Fast condition the threshold quickly decreases and is maintained at a very low level for the entirety of the task. The Careful-Fast condition has a relatively constant threshold until the change in incentive, after which the threshold sharply declines. However, the threshold in the Careful-Fast condition never reaches the low level of the Fast-Fast condition, even after 1000 decisions in the new incentive scheme. Comparison of the Careful-Careful and Fast-Careful conditions shows a corresponding pattern. The Careful-Careful condition has a relatively constant threshold throughout the task. In response to the change to the Careful incentive, the Fast-Careful threshold increases but it never reaches the level of the Careful-Careful threshold.

Table 8 shows the simulated median decision time, accuracy, and threshold of the last 500 decisions of each phase for each participant. The higher mean threshold of the Careful-Fast condition generated slower and more accurate decisions in stage 2 than the Fast-Fast condition. Similarly, the lower mean threshold of the Fast-Careful condition generated faster and less accurate decisions in stage 2 than the Careful-Careful condition.

The simulation provides a straightforward illustration of how perception shifts can

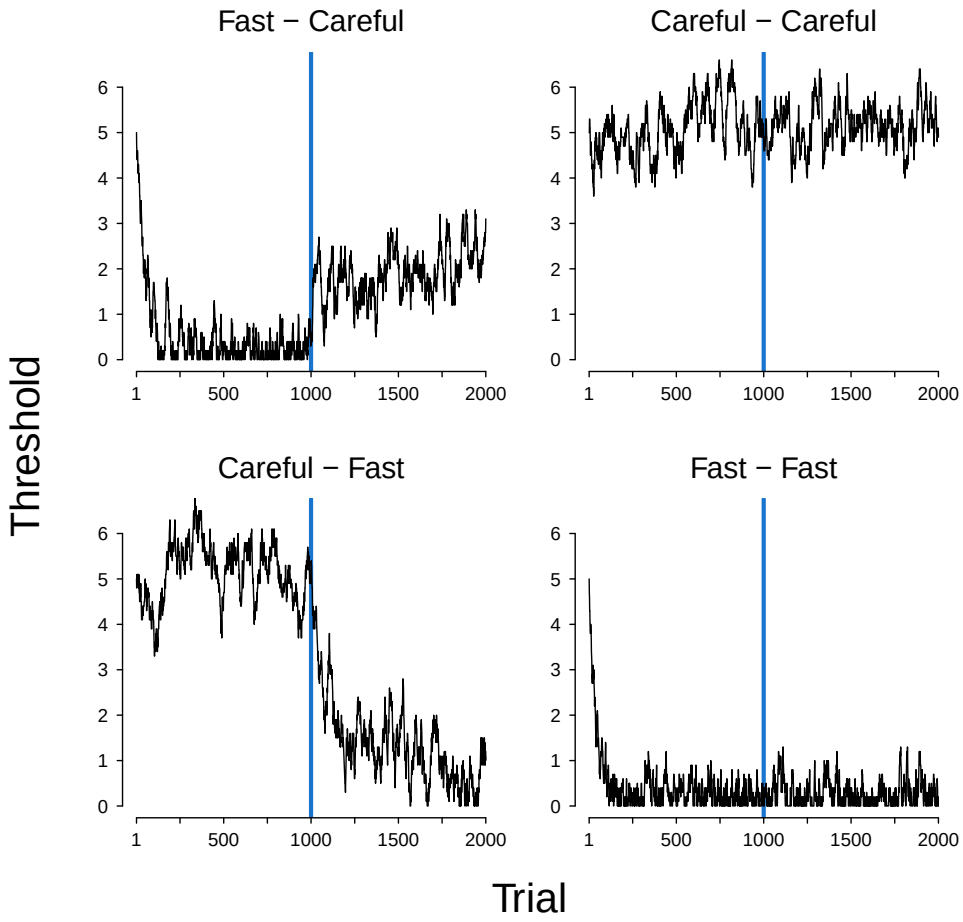

Figure 7. A Bayesian updating mechanism guides threshold adaptation in the four conditions of Experiment 1. See main text for details.

produce lasting differences in thresholds for people with identical incentives. When given an incentive to make slower and more accurate decisions, the Fast-Careful participant began increasing the threshold in response to every incorrect response. As the threshold increased and decisions became slower, virtually every decision was perceived to be slow, so the participant increased the threshold by only a small amount. The participant's perceptions of speed, based on extensive experience with fast decisions, resulted in the participant effectively desiring greater speed. Similarly, the Careful-Fast participant, upon decreasing the threshold to enable faster decisions, perceived every decision to be fast, and thus only modestly decreased the threshold after every correct decision.

Table 8
Simulation: Median response time, accuracy, and mean threshold of the last 500 decision in each condition and stage.

Condition	Stage	Median Decision Time (s)	Accuracy	Mean Threshold
F-C	1	2.23	.52	0.24
	2	3.31	.60	2.05
C-C	1	5.13	.69	5.32
	2	5.01	.70	5.17
C-F	1	5.03	.70	5.12
	2	2.77	.58	0.88
F-F	1	2.31	.52	0.27
	2	2.37	.52	0.26

General Discussion

In any choice situation, the decision maker must at some point determine that the choice is "made," or in other words that they have settled upon a preferred option or course of action. Only rarely can a decision maker reach 100% confidence in the superiority of a given choice option over all others, so a fundamental aspect of all decisions is the determination of the level of confidence that should be reached before a decision can be made or settled. This level of confidence, commonly referred to as a decision threshold, is the primary mechanism by which decision makers trade off the speed and accuracy of their decisions. A high threshold leads to greater accuracy but also to slower decisions, while a low threshold generates fast decisions that are more likely to be wrong. In repeated decisions, decision makers can quickly adjust their threshold based on the outcome of earlier decisions (Gold & Shadlen, 2007; Simen et al., 2006). If a decision takes too long, a decision maker can lower their threshold on the next decision occasion. If a decision results in a poor choice outcome, the decision maker is likely to increase their threshold on the next occasion. Such self-regulating models of decision thresholds exist in the literature (eg., Vickers & Lee, 1998, 2000).

The current research demonstrates two mechanisms that influence this learning process and can lead to biased thresholds. First, people's perceptions of *fast* and *slow*, *accurate* and *inaccurate* are context-specific, and thus these evaluations are based on a decision

maker's original experience in a decision environment. When decision makers adjust their thresholds, they may evaluate the resulting decision as being sufficiently fast or sufficiently accurate even when they are quite slow or quite inaccurate. These biased evaluations lead to biased thresholds. Second, decision environments can activate earlier decision goals to *be fast* or to *be accurate*. As a result, later decisions will be influenced by these earlier decision goals and thereby bias thresholds. Extant research on speed-accuracy tradeoffs has not recognized the potential for perception shifts and goal activation to affect thresholds. Both of these mechanisms, along with a third mechanism (decision strategy carryover), bias the overall threshold level in the same direction. In repeated decisions, thresholds are sticky—decision makers tend to use thresholds closer to their original threshold than decision makers who do not have this earlier decision experience. Four experiments showed evidence not only of the threshold bias, but also of the independent contribution of the two mechanisms (while controlling for decision strategy carryover).

We now address two alternate explanations of sticky thresholds. One alternative explanation is mindset carryover. Shen, Wyer Jr, and Cai (2012) found that performing a task quickly (vs. slowly) caused participants to perform a later, unrelated task more quickly (vs. more slowly) due to carryover of a "do things quickly" mindset. An important distinction made by Shen et al. is that their observed effect did not occur due to a carryover of a "do things quickly" goal, but due to a carryover of a mindset, which is a set of procedures enacted to pursue a goal (Gollwitzer, 2012). Because mindsets function at the procedural level, and not at the goal level, a "do things quickly" mindset would not affect the threshold itself but would rather change the procedures enacted to make the decision. The SAT models we fit to the data from Experiments 1 and 2 found empirical evidence against a mindset explanation. When we estimated a SAT model that allowed for different sensitivity parameters to detect differences in the pace of information extraction, this model provided a poorer explanation of the data relative to the model that allowed for different response threshold parameters. A "do things quickly" mindset could also be evidenced through a reduction in non-decision time. A model that allowed for different non-decision time

parameters also provided a relatively poorer explanation of the data. The outcome of our model comparison reduces the plausibility of mindset carryover being a cause of sticky thresholds.

A second potential alternative explanation of our results is anchoring and adjustment. This explanation posits that decision makers' original threshold serves as an anchor that results in underadjustment by the same mechanism that generates underadjustment in previous demonstrations of the effect of irrelevant numerical anchors (Tversky & Kahneman, 1974). This explanation cannot account for why thresholds are biased in repeated decisions with observable feedback. Anchoring and adjustment is typically observed in one-shot estimation tasks, not in repeated tasks. Our proposed mechanisms of perception shifts and goal activation provide a more coherent explanation because they can survive repeated feedback.

The three-fold contribution of this paper—the phenomenon of threshold stickiness and its two explanatory mechanisms—are applicable to repeated decisions. Given that many decisions faced by people are in some way repeated, the effect and its mechanisms constitute an important piece of knowledge for the SAT literature. But perhaps a more important contribution of this research is the added knowledge it generates about decision making generally. The fact that perceptions of speed and accuracy are context-specific and can thereby bias threshold levels has not been examined by prior research, and this influence could have a sizeable influence on all decisions, whether repeated or not. Consider a couple deciding where to honeymoon. Because of a strong desire to "get it right" in this first-time decision, both members of the couple are likely to set a relatively high threshold. But perceptions of speed are likely to be affected by previous vacation decisions. If one member of the couple is accustomed to doing extensive research, while the other member of the couple enjoys making vacation decisions with little forethought, their thresholds could be dramatically different, even with a similarly strong goal to be accurate in the decision. Future research could be directed at how decision makers form perceptions of speed and accuracy in first-time decisions, including features of the decision that affect

such perceptions as well as how perceptions in other decision contexts influence related contexts.

The findings outlined in this paper highlight the need for additional research on the influence of contextual factors on threshold levels. Decision environments can activate prior decision goals, which affects decisions in that particular decision environment, but that activation can also spread to other decisions. Many decisions share some contextual features, so any decision goals associated with a particular decision context may be activated in other decision contexts that share some of the same features. For example, all decisions within the same supermarket share many contextual features, so thresholds on all of the decisions within the store may be subject to a similar influence. Prior research on consumer decision making has shown that mobile users make systematically different decisions than desktop users (Ghose, Goldfarb, & Han, 2013). While this research inferred that differing search costs across the two device types accounted for these behavioral differences, our research indicates another possible cause. Mobile users, having grown accustomed to making faster decisions on their mobile devices, are likely to set lower thresholds for all decisions on their mobile devices, because their perceptions of decision speed are made in the context of mobile usage.

We hope this research spurs follow-up research not just in the area of biased thresholds but also in other areas of decision metacognition. By "decision metacognition", we refer to thoughts and beliefs about our own choice- and decision-directed cognition. Large interpersonal differences in decision time for a particular choice may result from differences in decision metacognition. Some decision makers may believe in their ability to think their way to the best choice, while others may believe that so many factors lie outside their control that choice-directed cognition is mostly wasted. Differences in metacognition may also help explain why some people rely on decision aides and expert opinions, while other consumers prefer to "go it alone" in their research. Examining the determinants of these beliefs and ways to influence them may be fruitful areas of future research.

References

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.

Bargh, J. A., & Barndollar, K. (1996). Automaticity in action: The unconscious as repository of chronic goals and motives.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American psychologist*, *54*(7), 462.

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.

Bogacz, R., Hu, P. T., Holmes, P. J., & Cohen, J. D. (2010). Do humans produce the speed–accuracy trade-off that maximizes reward rate? *The Quarterly Journal of Experimental Psychology*, *63*(5), 863–891.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in neurosciences*, *33*(1), 10–16.

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(4), 904.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive psychology*, *57*(3), 153–178. doi: 10.1016/j.cogpsych.2007.12.002

Busemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, *32*(2), 91–134.

Cooper, G., Innes, R., Kuhne, C., Cavallaro, J.-P., Gunawan, D., Hawkins, G., & Brown, S. (2021). pmwg: Particle metropolis within gibbs [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=pmwg` (R package version 0.2.0)

Dutilh, G., Annis, J., Brown, S. D., Cassey, P., Evans, N. J., Grasman, R. P., ... Donkin, C. (2019). The quality of response time data inference: A blinded, collaborative assessment of the validity of cognitive models. *Psychonomic Bulletin & Review*, *26*(4), 1051–1069. doi: 10.3758/s13423-017-1417-2

Evans, N. J. (2021). Think fast! the implications of emphasizing urgency in decision-making. *Cognition*, *214*, 104704.

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.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*(36), 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., 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 Sciences*, *105*(45), 17538–17542.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, *67*, 641–666.

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How is the mobile internet different? search costs and local activities. *Information Systems Research*, *24*(3), 613–631.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, *30*.

Gollwitzer, P. (2012). *Mindset theory of action phases*.

Gunawan, D., Hawkins, G. E., Tran, M.-N., Kohn, R., & Brown, S. (2020). New estimation approaches for the hierarchical linear ballistic accumulator model. *Journal of Mathematical Psychology*, *96*, 102368.

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012a). Context effects in multi-alternative decision making: empirical data and a bayesian model. *Cognitive Science*, *36*(3), 498–516.

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012b). 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. *Journal of Neuroscience*, *35*(6), 2476–2484.

Hawkins, G. E., Hayes, B. K., & Heit, E. (2016). A dynamic model of reasoning and memory.

Journal of Experimental Psychology: General, *145*(2), 155.

Hawkins, G. E., & Heathcote, A. (2021). Racing against the clock: Evidence-based versus time-based decisions. *Psychological Review*, *128*(2), 222–263.

Hawkins, G. E., Marley, A., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014a). The best of times and the worst of times are interchangeable. *Decision*, *1*(3), 192.

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

Heathcote, A., Brown, S. D., & Wagenmakers, E.-J. (2015). An introduction to good practices in cognitive modeling. In *An introduction to model-based cognitive neuroscience* (pp. 25–48). Springer.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150.

Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, *76*(3), 616–628.

Helson, H. (1964). Adaptation-level theory: an experimental and systematic approach to behavior.

Huang, A., & Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439–452.

Jeffreys, H. (1967). Theory of probability. (corrected). *New York*.

Kalbfleisch, J., & Prentice, R. (1980). Failure time methods. In *The statistical analysis of failure time data* (pp. 13–14). John Wiley, New York.

Lerche, V., & Voss, A. (2017). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological research*, 1–16.

Levav, J., Reinholtz, N., & Lin, C. (2012). The effect of ordering decisions by choice-set size on consumer search. *Journal of Consumer Research*, *39*(3), 585–599.

Mittner, M., Boekel, W., Tucker, A. M., Turner, B. M., Heathcote, A., & Forstmann, B. U. (2014). When the brain takes a break: a model-based analysis of mind wandering. *Journal of Neuroscience*, *34*(49), 16286–16295.

Myung, I. J., & Busemeyer, J. R. (1989). Criterion learning in a deferred decision-making task. *The American journal of psychology*, 1–16.

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*,

25(4), 1225–1248.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between stroop and simon effects using delta plots. *Attention, Perception, & Psychophysics*, *72*(7), 2013–2025.

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

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, *111*(2), 333.

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.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological review*, *106*(2), 261.

Rieskamp, J., & Otto, P. E. (2006). Ssl: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review*, *108*(2), 370.

Shen, H., Wyer Jr, R. S., & Cai, F. (2012). The generalization of deliberative and automatic behavior: The role of procedural knowledge and affective reactions. *Journal of Experimental Social Psychology*, *48*(4), 819–828.

Simen, P., Cohen, J. D., & Holmes, P. (2006). Rapid decision threshold modulation by reward rate in a neural network. *Neural networks*, *19*(8), 1013–1026.

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.

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

Stevens, S. S. (2017). *Psychophysics: Introduction to its perceptual, neural and social prospects*. Routledge.

Treisman, M., & Williams, T. C. (1984). A theory of criterion setting with an application to sequential dependencies. *Psychological review*, *91*(1), 68.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion models of decision–making. In V. M. Sloutsky, B. C. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1429–1434). Cognitive Science Society.

Van Vugt, M. K., Simen, P., Nystrom, L., Holmes, P., & Cohen, J. D. (2014). Lateralized readiness potentials reveal properties of a neural mechanism for implementing a decision threshold. *PloS one*, *9*(3).

Verdonck, S., & Tuerlinckx, F. (2014). The ising decision maker: A binary stochastic network for choice response time. *Psychological Review*, *121*(3), 422–462.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. properties of a self-regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*(3), 169–194.

Vickers, D., & Lee, M. D. (2000). Dynamic models of simple judgments: Ii. properties of a self-organizing pagan (parallel, adaptive, generalized accumulator network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences*, *4*(1), 1–31.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental psychology*, *60*(6), 385.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & cognition*, *32*(7), 1206–1220.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The*

Annals of Mathematical Statistics, 326–339.

Wall, L., Gunawan, D., Brown, S. D., Tran, M.-N., Kohn, R., & Hawkins, G. E. (2021). Identifying relationships between cognitive processes across tasks, contexts, and time. *Behavior Research Methods*, *53*(1), 78–95.

Winkel, J., Hawkins, G. E., Ivry, R. B., Brown, S. D., Cools, R., & Forstmann, B. U. (2016). Focal striatum lesions impair cautiousness in humans. *cortex*, *85*, 37–45.

Zacksenhouse, M., Bogacz, R., & Holmes, P. (2010). Robust versus optimal strategies for two-alternative forced choice tasks. *Journal of mathematical psychology*, *54*(2), 230–246.

Appendix A

Response Distributions

Figure A1 summarizes group-level and individual-participant data from Experiments 1 and 2. Measures of central tendency (mean – Tables 1 and 4 of the main text; median – Figure A1) provide a reasonable summary of the condition-level data. The spread of participant summary statistics is reasonably even across conditions, with some minor suggestion of greater individual differences in some conditions (e.g., mean response time for the C-F condition in Experiment 1) relative to others (e.g., mean response time for the F-C condition in Experiment 1).

Appendix B

Parameter Recovery

We performed a parameter recovery study to confirm the parameters of the LBA model of the pricing task in Experiment 1 are reliably recovered from data. The study was based on the design of Experiment 1 and was performed independently for each of the four models (null, threshold, utility, non-decision time). For each model, we first randomly sampled a parameter vector from each participant's posterior distribution of the LBA parameters estimated from the real data. We used these parameter vectors to generate a synthetic data set of the same size as the real data; the number of trials each participant completed in each stage, and the number of participants in each of four conditions (F-C, C-C, C-F, F-F). We then estimated the parameters of the LBA model from these simulated data using identical methods to those applied to the real data. We independently repeated the procedure 10 times for each model. This produced 10 independent hierarchical estimation exercises that assessed parameter recovery from a total of 760 participant-level posterior distributions for each of the four models. The focus of the recovery analysis was at the participant level.

Figure B1 shows the results of the parameter recovery study for the threshold model (i.e., the DIC-preferred model). Overall, parameters of the model were well recovered, in-

dicated by the points largely falling close to the identity lines. Parameter recovery results for the remaining three models are shown in Figure B2 (sensitivity model), Figure B3 (non-decision time model), and Figure B4 (null model). In each case there is good parameter recovery. One exception appears to be the sensitivity parameter for some models (e.g., Figure B4). However, this is due to the scale of the transformation: a probit transformation of values >2 lead to approximately the same value on the raw scale. These results demonstrate that the parameters of the LBA model can be recovered in typically-sized data sets in the pricing task.

Appendix C

Experiment 3: Additional Analysis

Table C1 provides participants' mean response times by condition, by portion of the experiment, and by decision type—Easy or Hard. During the first part of the experiment, participants in the Fast condition made Easy decisions with roughly the same speed as participants in the Careful condition and they made Hard decisions more quickly. However, because participants in the Careful condition faced a higher proportion of Easy questions, their overall decision speeds were identical in this part of the experiment, as described in the main text. In the second part of the experiment, the speed goal of participants in the Fast condition is activated by the decision context and leads to faster decision speeds, as reported in the main text.

Table C1
Experiment 3: Average response time by condition, decision type, and portion of the experiment (in seconds).

Condition	Decision type	Part 1	Part 2
Fast	Easy	5.54	4.82
	Hard	7.18	6.39
Careful	Easy	5.51	5.18
	Hard	7.87	7.47

The faster decision speeds by participants in the Fast condition would typically lead to lower decision accuracy. Table C2 shows participants' mean accuracy by condition, portion of the experiment, and decision type. Oddly, participants in the Fast condition were more accurate in their decisions in the second part of the experiment, contrary to expectations. This difference in accuracy was not significant for either the Easy questions ($z = 1.64$, $p = .102$) or for the Hard questions ($z = .653$, $p = .514$), so the data do not indicate superior performance for the Fast participants over the Careful participants, but the means in the wrong direction are nevertheless puzzling. We speculate that because Fast participants saw a greater number of Hard questions in the first part of the experiment, they were more practiced at answering these more difficult questions, so their faster speeds in the second part of the experiment did not produce lower accuracy levels. Meanwhile,

the Careful participants, who were more practiced at the Easy questions, did not exhibit higher accuracy levels in these questions because of a ceiling effect. Accuracy was nearly perfect on the Easy questions in both conditions, so we observed no difference in accuracy. Nevertheless, since there was not a significant difference in accuracy between the Fast and Careful conditions for either Easy or Hard decisions, the accuracy results do not provide evidence contrary to our key hypothesis.

Table C2
Experiment 3: Accuracy by condition, decision type, and portion of the experiment (in percent).

Condition	Decision type	Part 1	Part 2
Fast	Easy	91.2	98.5
	Hard	74.6	76.5
Careful	Easy	95.3	96.5
	Hard	73.3	72.4

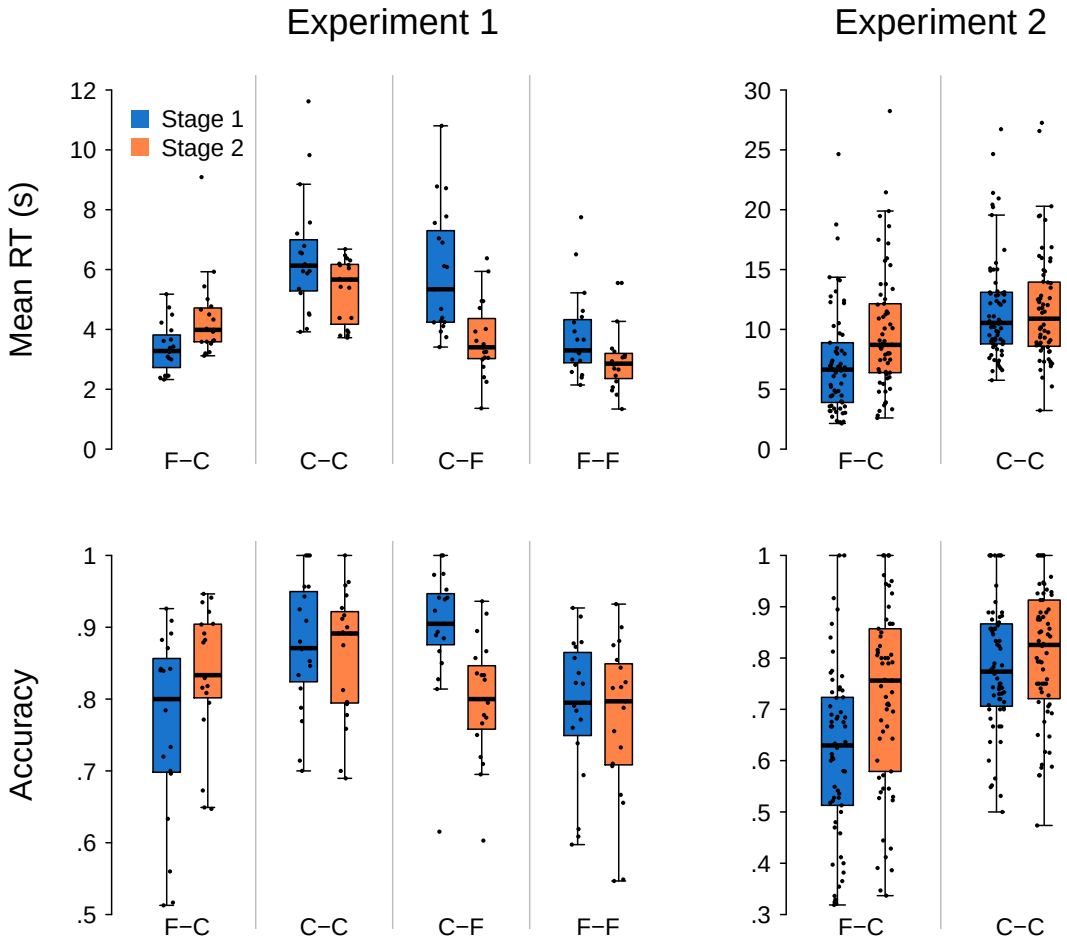

Figure A1. Condition-level and individual-participant data summaries for Experiments 1 and 2 (columns). Mean response time and accuracy are shown in rows. Within panels, x-axis position displays experimental condition (F-C, C-C, C-F, F-F) and color indicates stage of the experiment (1, 2). Condition-level summaries are shown with boxplots and individual-participant data are shown with dots.

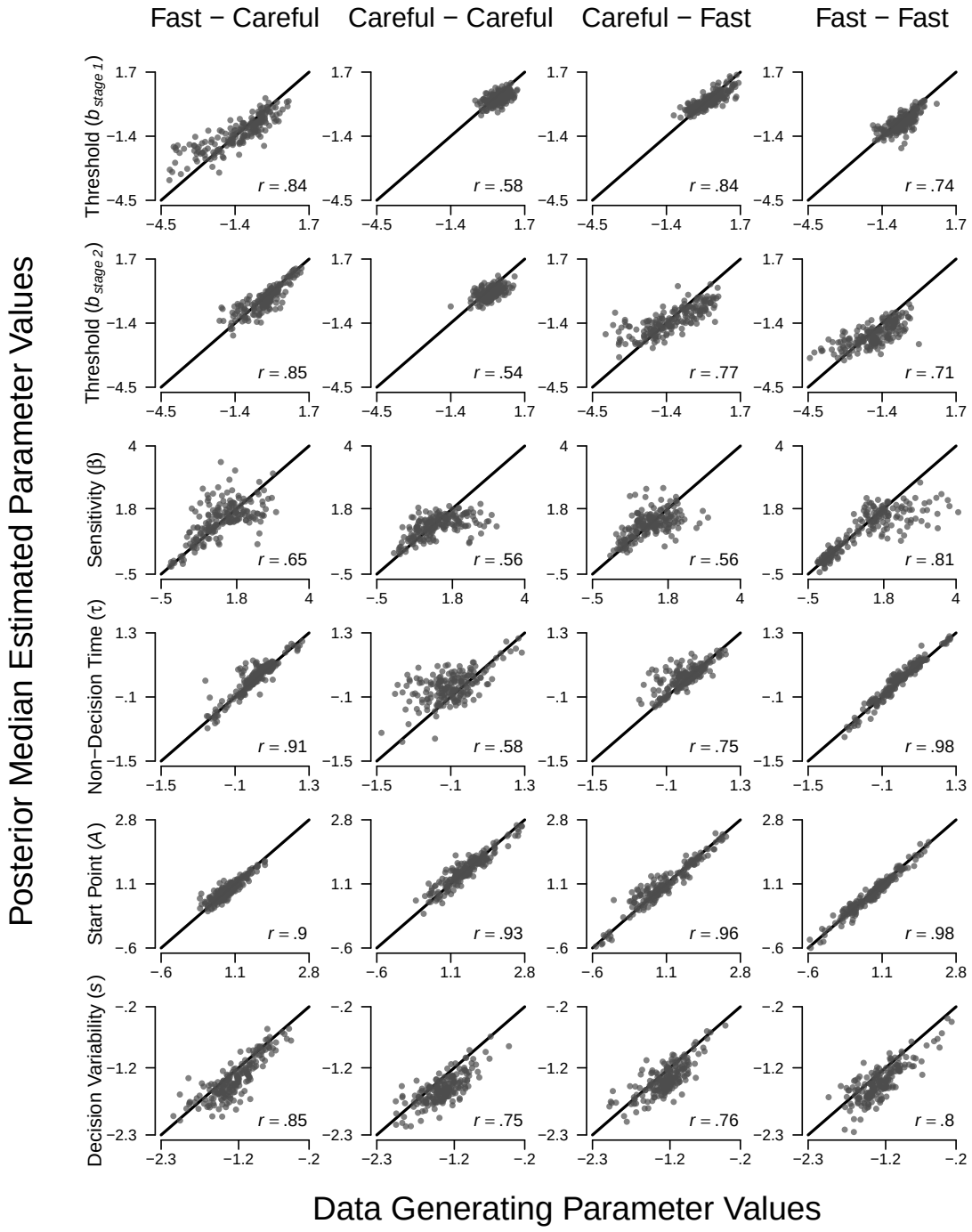

Figure B1. Parameter recovery for the threshold model applied to the pricing task of Experiment 1. Model parameters are shown in rows and conditions of the experiment are shown in columns. Within each panel, data-generating parameter values are shown on the x-axes and the median of the estimated posterior distributions are shown on the y-axes. Axis scaling is constant across conditions for each parameter to facilitate comparison. Perfect parameter recovery is indicated along the diagonal. Parameters are shown in the transformed space in which they were estimated (probit scaling for sensitivity, log scaling for all other parameters).

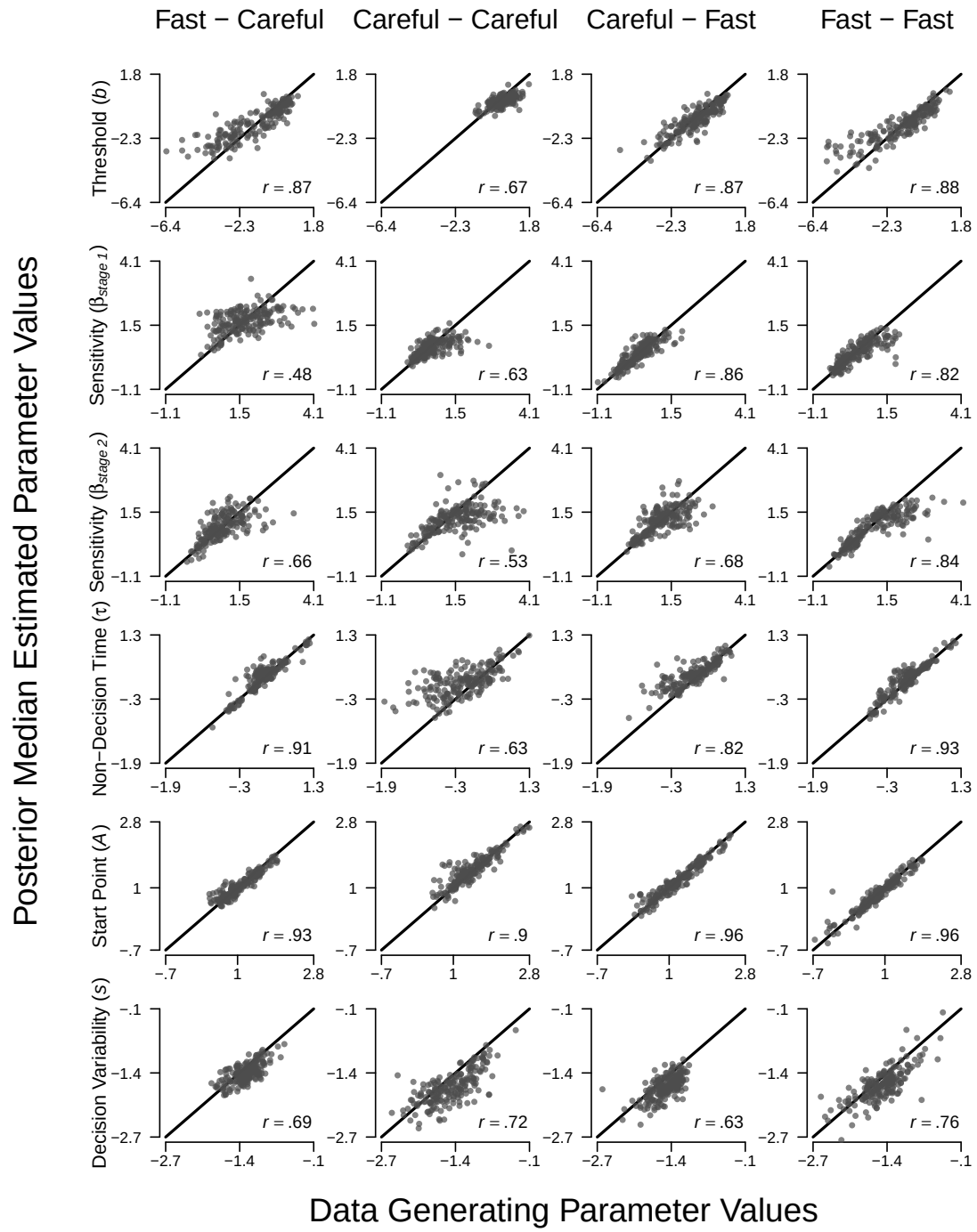

Figure B2. Parameter recovery for the sensitivity model applied to the pricing task of Experiment 1. Details are as described in Figure B1.

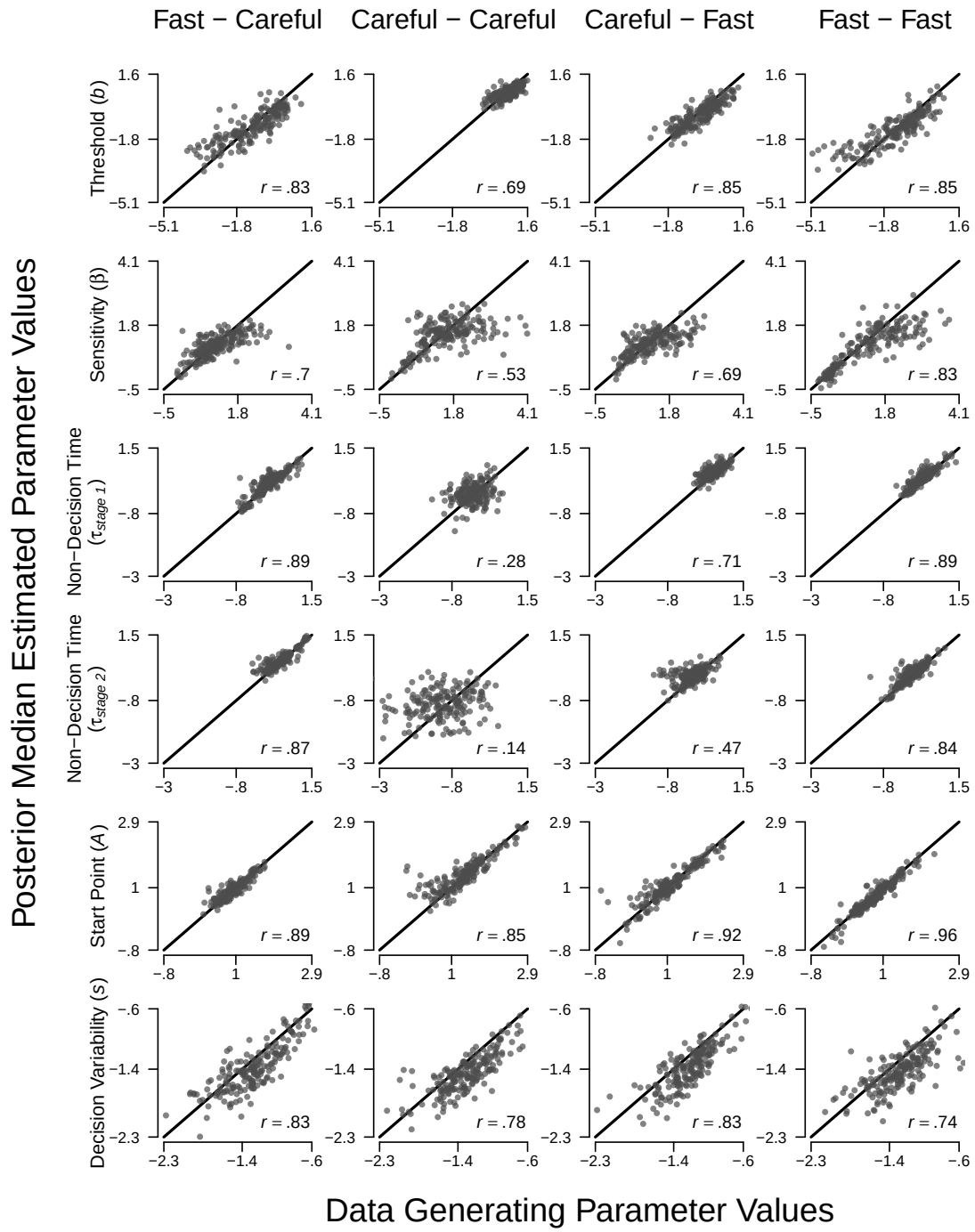

Figure B3. Parameter recovery for the non-decision time model applied to the pricing task of Experiment 1. Details are as described in Figure B1.

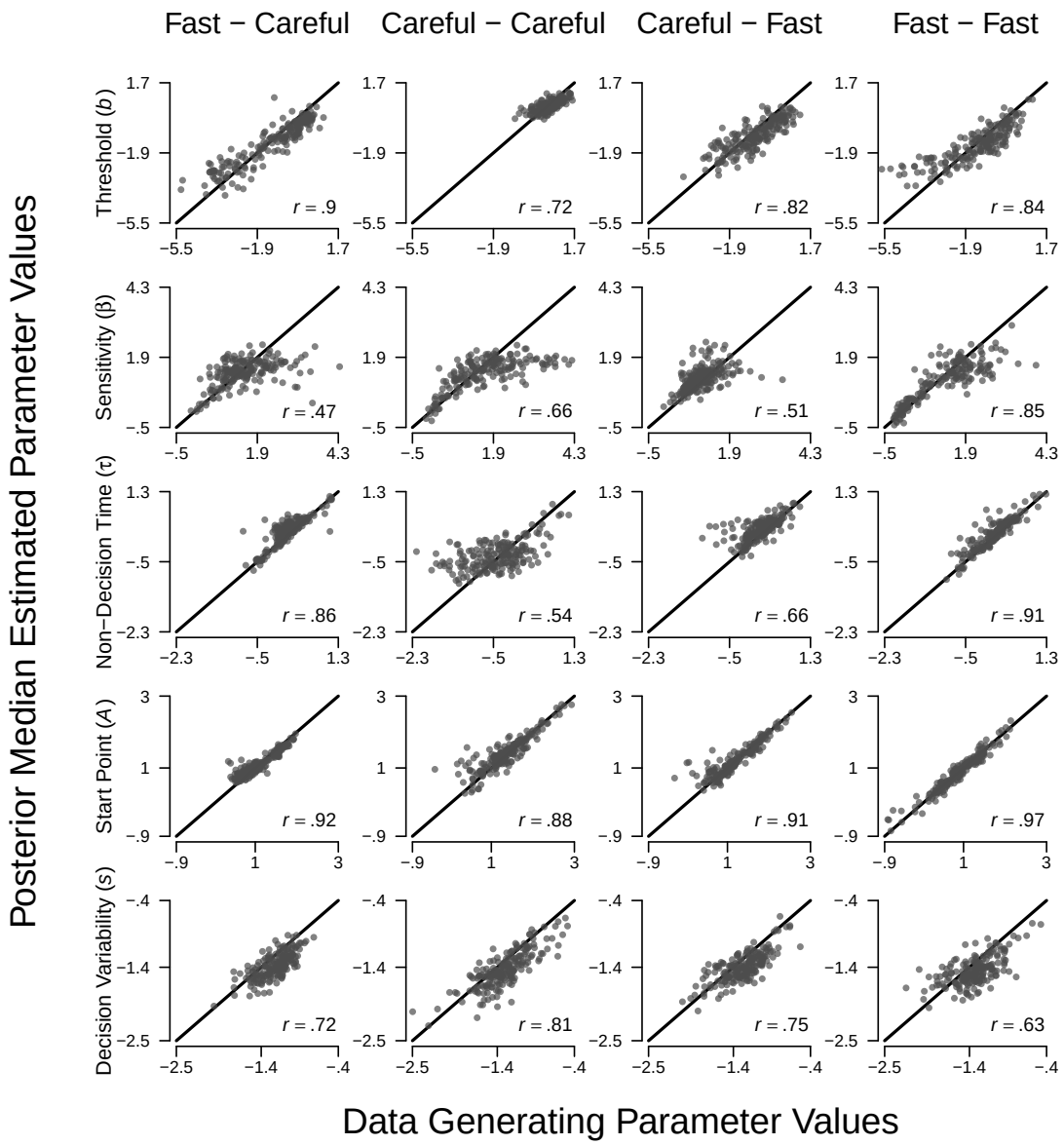

Figure B4. Parameter recovery for the null model applied to the pricing task of Experiment 1. Details are as described in Figure B1.