Humans reconfigure target and distractor processing to address distinct task demands

Harrison Ritz[*1-3] & Amitai Shenhav[1,2]

1. Cognitive, Linguistic & Psychological Science, Brown University, Providence, RI, USA
2. Carney Institute for Brain Science, Brown University, Providence, RI, USA
3. Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

** Corresponding author:* hritz@princeton.edu

Acknowledgements: This work was supported by the Daniel Cooper Graduate Student Fellowship (H.R.), as well as grants R01MH124849 and NSF CAREER Award 2046111 (A.S.). We are grateful to Kia Sadahiro, William McNelis, Allison Loynd, and Savannah Doelfel for assistance in data collection, and to Michael J. Frank, Matthew N. Nassar, Jonathan Cohen, David Badre, Tobias Egner, Senne Braem, Sebastian Musslick, and both the Shenhav Lab and LNCC for helpful discussions on these topics.

Conflicts of Interest: None

Abstract

When faced with distraction, we can focus more on goal-relevant information (targets) or focus less on goal-conflicting information (distractors). How people use cognitive control to distribute attention across targets and distractors remains unclear. We address this question by developing a novel Parametric Attentional Control Task (PACT) that can 'tag' participants' sensitivity to target and distractor information. We use these precise measures of attention to develop a novel process model that can explain how participant control attention towards target and distractors. Across three experiments, we find that participants met the demands of this task by independently controlling their processing of target and distractor information, exhibiting distinct adaptations to manipulations of incentives and conflict. Whereas incentives preferentially led to target enhancement, conflict on the previous trial preferentially led to distractor suppression. These distinct drivers of control altered sensitivity to targets and distractors early in the trial, promptly followed by reactive reconfiguration towards task-appropriate feature sensitivity. To provide a process-level account of these empirical findings, we develop a novel neural network model of evidence accumulation with attractor dynamics over feature weights that reconfigures target and distractor processing. These results provide a computational account of control reconfiguration that provides new insights into how multivariate attentional signals are optimized to achieve task goals.

Keywords: Cognitive Control; Decision-Making; Attention; Evidence Accumulation

Introduction

Whether we are having a conversation in a crowded coffee shop or writing a paper at our desk while surrounded by browser tabs, most tasks require us to engage in two distinct forms of attentional control[1]. One form of control enhances the processing of task-*relevant* information, for instance by paying careful attention to what our conversation partner is sharing with us. The other form of control suppresses the processing of task-*irrelevant* information, particularly that which conflicts with our primary goal (e.g., distraction from a nearby conversation). While past research has extensively studied target and distractor processing, it has done so primarily by focusing on each one separately. As a result, relatively little is known about how control over task-relevant information (targets) interacts with control over task-irrelevant information (distractors). Can people control multiple forms of information processing, and if so, how do they regulate these information streams over time? Here, we bridge previous methodological gaps to gain new insight into the top-down control over target and distractor processing, providing an integrative model of how dynamic control adjustments could occur within and across trials.

Research into how people enhance the target of their attention versus actively suppress distractors has been largely governed by separate research areas, using different approaches. Studies of perceptual decision-making have characterized the process by which people try to

[1] Through-out, we refer to 'cognitive control' as the process that configures information processing to achieve task goals (Botvinick and Cohen, 2014). Whereas cognitive control refers to all such adjustments, such as changes to stimulus sensitivity or decision threshold, we reserve 'attentional control' for just the top-down control over stimulus sensitivity.

determine the correct response (e.g., which of two categories this stimulus belongs to) based on noisy information about a target stimulus, and how this varies with the difficulty of discriminating that stimulus (e.g., how perceptually similar two stimuli are; (Britten et al., 1992; Gold and Shadlen, 2007). This contrasts with studies of inhibitory control, in which the correct response to a target is typically unambiguous (e.g., respond left when seeing a high-contrast leftward-facing arrow), but a second dimension of the stimulus display (one that is typically processed more automatically; e.g., flanking arrows pointing rightward) triggers a conflicting response (Botvinick and Cohen, 2014; Posner and Snyder, 1975).

Despite the substantial progress that has been made in understanding these two processes in parallel, critical questions remain that can only be addressed by studying them in tandem (Ritz et al., 2022). Most notably, it is unclear how people decide how to distribute their control between targets and distractors. When the demands or incentives for performing a task change, do people re-direct control towards target enhancement, distractor suppression, or both? For instance, previous work has shown that people are less susceptible to the influence of distractors after overcoming a previously conflicting distractor (the so-called *conflict adaptation* or *congruency sequence* effect; (Gratton et al., 1992). Prevailing models have accounted for these findings by assuming that participants increase attention to the target dimension following a high-conflict trial (Botvinick et al., 2001; Egner, 2007), but limitations of the relevant experiments (e.g., most experiments don't manipulate target salience; through see (Lindsay and Jacoby, 1994; Servant et al., 2014; Stafford et al., 2011)) make it difficult to rule out that adaptation may also occur at the level of distractor suppression (Lindsay and Jacoby, 1994; Tzelgov et al., 1992). It is more generally an open question whether effects of recent task difficulty (e.g., low discriminability or

high-conflict) result in control-specific or control-general adaptations and, similarly, whether the motivation to improve performance in such settings leads to preferential engagement of one or both forms of control. Cognitive control is fundamentally an adaptive process, so people's specific control policies should depend on the task structure (Botvinick and Cohen, 2014; Egner, 2008). However, understanding how people coordinate multiple forms of information processing can help inform the architecture of the underlying control process (Ritz et al., 2022).

One way that previous research has studied target and distractor adjustments is to measure changes in brain activity associated with task-relevant stimulus processing. For example, some previous work has suggested that conflict-triggered control preferentially enhances sensitivity in regions associated with target stimuli (e.g., faces in fusiform face gyrus (Egner and Hirsch, 2005). Other studies have found evidence for both target and distractor control by using similar stimulus-tagging methods (Gazzaley et al., 2005; Soutschek et al., 2015) or by exploiting lateralized EEG responses (Noonan et al., 2016; Wöstmann et al., 2019). The range of results across these neuroimaging experiments may come from the different tasks that have been deployed (Egner, 2008), but may also arise from noisy or complex correspondence between neuroimaging methods and underlying cognitive processes. In the current experiment, we provide new methods for indexing target and distractor sensitivity from behavior alone, enabling us to provide new insight into the cognitive architecture of feature-selective control.

Recent models of controlled decision-making have emphasized the role that within-trial attentional dynamics play in response conflict tasks, offering new insight into the implementation of cognitive control (Servant et al., 2014; Weichart et al., 2020; White et al.,

2011; Yu et al., 2009). These models have largely focused on the Eriksen flanker task, modelling how an attentional spotlight centered on the target item narrows over time. This formulation necessarily yokes target enhancement and distractor suppression due to the spatial spread of attention. As a result, little is known about whether target and distractor processing dynamics can fall under independent control when these are not explicitly yoked, as in the case of feature-based attention. Less still is known about how adjustments driven by factors like conflict adaptation and incentives alter the *dynamics* of target and distractor processing (Adkins and Lee, 2021).

To address these questions, we developed a novel task that orthogonally varies target and distractor information, measuring how processing of these two dimensions varies both within and across trials. Our task merges elements of paradigms that have been separately popularized within the two research areas above. To capture variability in target processing, we based our task on the random dot kinematogram paradigm (Danielmeier et al., 2011; Kang et al., 2021; Kayser et al., 2010; Mante et al., 2013; Shenhav et al., 2018). This task parametrically varies the motion discriminability (e.g., percentage of dots moving left) and color discriminability (e.g., percentage of green dots) across an array of dots. Participants were instructed to respond to the color dimension, while ignoring the motion dimension. Critically, whereas color response mappings were arbitrary (e.g., left hand for green), motion responses were exactly aligned with the direction of motion (e.g., left hand for leftward moving stimuli), resulting in potent "Simon-like" [2] response interference from this prepotent distractor. A salient incongruent distractor

[2] The Simon task is a classic cognitive control task in which participants must ignore a response-affording stimulus feature (e.g., respond 'left' to a leftwards spatial location), and instead respond to a less prepotent stimulus feature (e.g., respond 'right' to a blue stimulus). The classic pattern of results is that participants perform more poorly when these two features disagree than when they correspond to the same response (see: (Egner, 2007)).

provokes an erroneous response, providing a qualitatively different form of difficulty from how low coherence targets make it harder to choose the correct answer (Norman and Bobrow, 1975).

Previous work has demonstrated response conflict and trial-to-trial adjustments in a color-motion kinematogram with full target coherence and binary distractor congruence (Danielmeier et al., 2011). We extended this task by parametrizing both target coherence and distractor congruence. In doing so, we are able to obtain more precise measures of feature sensitivity by accounting for global performance factors (e.g., lapse rate; (Wichmann and Hill, 2001). Importantly, however, we can also isolate how participants simultaneously configure attention towards each of these feature dimensions. Using standard elicitors of cognitive control, namely performance-contingent incentives and response conflict, we examine how people dynamically configure both target and distractor gain to maximize their performance. We then use the precision afforded by these methodological advances to inform an explicit process model of attentional control.

We find that participants independently and dynamically control target and distractor processing over the course of a trial. To meet the demands of this task, participants preferentially enhanced target sensitivity under incentives, and preferentially suppressed distractor sensitivity after high conflict trials. Moreover, they implement these control strategies by changing the initial conditions of a dynamic process that enhances task-relevant feature processing and suppresses task-irrelevant feature processing. Finally, we find that these control strategies can be captured by extending classic neural network models of cognitive control to incorporate an attractor network that dynamically regulates the influence of different task features on choice. Together, these results extend our understanding of both decision-making and cognitive control by bridging

the methodological and theoretical divides between these fields, providing new insight into how we control multiple forms of information processing.

Methods

Participants

All participants provided informed consent in compliance with Brown University's Institutional Review Board, participating for either course credit or pay. We excluded participants from our analyses if they had <70% accuracy during attend-color blocks or completed less than half of the experiment. Fifty-seven individuals participated in Experiment 1 (mean(SD) age: 20.6(2.21); 36 female; 1 excluded), 42 individuals participated in Experiment 2 (age: 19.1(0.971); 31 female; 2 excluded), and 62 individuals participated in Experiment 3 (age: 19.8(1.38); 47 female; 2 excluded), resulting in 156 included participants across the three experiments. Sample sizes were guided by piloting in Experiment 1 and experimental standards in cognitive control research (commonly n = 20-40; e.g., (Adkins and Lee, 2021; Danielmeier et al., 2011; Jiang et al., 2015; Vogel et al., 2020; White et al., 2011).

Parametric Attentional Control Task (PACT)

We developed the Parametric Attentional Control Task (PACT), extending tasks used to study decision-making (Kang et al., 2021; Mante et al., 2013; Shenhav et al., 2018) and cognitive control (Danielmeier et al., 2011). On each trial, participants viewed an array of moving dots (i.e., random dot kinematogram), presented in one of four colors (see Figure 1). Participants were taught to match two colors to a left keypress and two colors to a right keypress (with colors

counterbalanced across participants). The majority color did not repeat on adjacent trials to avoid priming (Braem et al., 2019).

The direction of the dot motion (leftward or rightward) was task-irrelevant and could be consistent with the color response (distractor congruent trials) or it could be inconsistent with this response (distractor incongruent trials). Uniquely in this experiment, we parametrically varied the degree of distractor congruence on each trial by varying the motion coherence (percentage of dots moving in the same direction vs moving in a random direction). Distractor congruence was linearly spaced between 95% congruence and 95% incongruence, sampled randomly across trials. For variants with 11 levels of congruence, the congruence levels were [-95, -76, -57, -38, -19, 0, 19, 38, 57, 76, 95], with negative values being incongruent and positive values being congruent. We made the motion highly salient to maximize the conflict induced by this distracting dimension (Wöstmann et al., 2021), with dots moving quickly across a large aperture.

In Experiment 1, all of the dots were the same color (100% color coherence), creating a parametric extension of the Simon conflict tasks (Danielmeier et al., 2011). In Experiments 2 and 3, the dots contained a mixture of two colors associated with different responses. Color coherence was linearly spaced between 65% to 95%, drawn randomly across trials.

To maintain the salience of the motion dimension throughout the session (Shiffrin and Schneider, 1977), participants alternated between blocks of the task above (‘attend-color’ trials, putatively more control-demanding) and blocks where participants were instructed to instead indicate the

direction of the dot motion ('attend-motion' trials, putatively less control-demanding). Mirroring the attend-color blocks, in Experiment 1 we held the motion coherence constant (maximal) during attend-motion blocks, while varying the color coherence across trials. In Experiments 2 and 3, we varied the coherence of both dimensions during attend-motion blocks. In Attend-Motion trials, we allowed distractor colors to repeat on consecutive trials, mirroring the stimulus-repetitions that occurred in Attend-Color blocks.

Comparing performance across tasks that are matched for visual and motoric demands also allows us to test whether behavioral effects depend on stimulus or response confounds. For example, participants' behavior may be influenced by eye movement confounds (e.g., bottom-up attentional capture by motion coherence), response repetition biases (e.g., due to responses switching more often than repeating), or stimulus-response priming (e.g., due to how response switching coincides with stimulus transitions). Critically, Attend-Color and Attend-Motion tasks differ in their putative control demands, allowing us to isolate stimulus-response confounds from goal-directed control.

Session

Participants first performed 100 motion-only training trials (0% coherent color) and 100 color-only training trials (0% coherent motion; order counterbalanced across participants) to learn the stimulus-response mappings. During training, participants received accuracy feedback on every trial. During the main experiment, participants performed two types of interleaved blocks, without trial-wise feedback. Participants alternated between longer attend-color blocks (100 trials) and shorter attend-motion blocks (Experiment 1: 20-50 trials; Experiment 2-3: 30 trials; order counterbalanced across participants). In Experiments 1 and 2, at the end of each block

participants were told their average accuracy and median RT, and encouraged to respond quickly and accurately. Participants were not given this information in Experiment 3 to avoid interactions with the incentive manipulation (see below). Participants took self-timed breaks between blocks.

Stimuli

Participants were seated approximately 60cm from a computer screen, making their responses on a customizable gaming keyboard in a dark testing booth. The random dot motion array was presented in the center of the screen (~15 visual degrees in diameter; ~66.8 dots per visual degree squared; 19" LCD display at 60Hz). The dots colors were approximately (uncalibrated) isoluminant and perceptually equidistant (RGB: [187, 165, 222], [150, 180, 198], [192, 169, 168], [157, 184, 130]; (Teufel and Wehrhahn, 2000) and moved at ~15 visual degrees per second. Each trial started with a random inter-trial interval (Experiment 1: 0.5 – 1.5s; Experiment 2-3: 0.5 – 1.0s). There was an alerting cue 300ms before the trial onset, indicated by the fixation cross turning from grey to white, to minimize non-decision time. The stimuli were then presented until either a response was made, or a deadline was reached (Experiment 1: 3s; Experiment 2-3: 5s).

Task Variants

Experiment 1: These data incorporate several similar versions of this task developed during piloting. The main differences across versions were the number of distractor congruence levels (mean(range) = 13.5(11-15)), the number of trials per attend-motion block (mean(range) = 26(20-50)), and the total number of trials (mean(range) = 469(300-700) attend-color trials). We did not find significant differences in performance across versions, and so our analyses collapsed across these versions. Experiment 1 also included a learning condition in a separate set of blocks,

which was outside the scope of the current paper and not included in the analyses we report.

Experiments 2 & 3: These data come from a single task variant (though see Experiment 3's incentive manipulation below). In this variant, we presented participants with 11 levels of target coherence and 11 levels of distractor congruence, linearly spaced within their coherence range and randomly sampled across trials. Participants performed 12 blocks of 100 attend-color trials interleaved with 12 blocks of 30 attend-motion trials. Illustrative task instructions are provided in Supplementary Note 1.

Incentivized Variant (Experiment 3)

In Experiment 3, we studied task performance under monetary incentives to provide a convergent measure of control adjustments, to test where task processing was limited by motivation rather than hard constraints like stimulus information (Norman and Bobrow, 1975). We informed participants before the main session that they would be able to earn a monetary reward for good performance. On 'Reward' blocks, we randomly selected trials at the end of the experiment, and participants earned bonus payment for trials on which they were both fast (<75% of their RT distribution) and accurate. On 'No Reward' blocks, participants would not be eligible to earn a reward, but were encouraged to be fast and accurate. We indicated the incentive condition at the beginning of each block with a label and text coloring (gold text for 'Reward', white text for 'No Reward'). Participants were not instructed on the reward algorithm, only that they would earn rewards from being fast and accurate on randomly selected trials. Participants were not informed which trials were selected to avoid biasing post-reward trials. Participants performed Attend-Color and Attend-Motion blocks in one incentive condition before alternating to the other incentive condition (order counterbalanced across participants). At the end of the

experiment, participants received a bonus calculated from their performance (mean(SD) bonus: $2.5($0.57)USD).

Regression Analyses

We used a hierarchical nonlinear regression of choice and reaction time as a tractable and minimally theory-laden measure of performance (Supplementary Figure 1). We designed these regression models to quantify changes to target and distractor sensitivity, while controlling for global factors like behavioral autocorrection and how task factors may change lapse rates. The results of these regression analyses then provided the basis for our explicit process modeling (see below). We confirmed that our regression models are identifiable using Belsley collinearity diagnostics (collintest in MATLAB; Supplementary Table 10).

In particular, we implemented hierarchical expectation maximization (EM) in MATLAB R2020a (using emfit; available at github.com/mpc-ucl/emfit) to provide a maximum a posteriori (MAP) estimates for the mean and covariance of parameters linking task features to participants' reaction time and accuracy. This fitting algorithm alternates between finding the MAP estimates of participants' parameters given the current group-level expectations (M-step; with 5 parameter re-initialization per step), and updating this group-level expectation based on participants' estimated parameters (E-step), repeated until convergence. We fit separate regression to each experiment for independent replications of our findings. Analysis code is available at github.com/shenhavlab/PACT-public.

Our regression approach simultaneously estimated parameters for choice and RT:

$$logPost = logLike(Choice) + logLike(RT) + logPrior(Choice, RT)$$

Our choice sub-function used a lapse-logistic likelihood function, as previous work has shown that un-modelled lapse rates can mimic changes in psychometric slope (Wichmann and Hill, 2001). Our choice sub-function had the form:

$$Choice \sim \frac{1 - lapse}{1 + exp(-\beta_{Choice} X_{Choice})} + (lapse \times 0.5)$$

$$lapse = \frac{1}{1 + exp(-\beta_{Lapse} X_{Lapse})}$$

Where β_{Choice} and β_{Lapse} are parameter vectors, and X_{Choice} and X_{Lapse} are design matrices. Our RT sub-function used a shifted lognormal likelihood function:

$$log(RT - ndt) \sim \beta_{RT} X_{RT}$$

Where again $\beta_{RT} X_{RT}$ is a linear model, and ndt is the estimated non-decision time. Rare RTs less than ndt were assigned a small likelihood. This helped avoid one fast RT from unduly influencing this parameter, while still capturing these informative trials.

Finally, the prior probability of the parameters was evaluated under a multivariate normal distribution defined by the group-level parameter mean and covariance, improving the robustness of our estimates through regularization. Critically, we estimated this group-level covariance

across both choice and RT parameters, which better regularized our estimates and produced a joint model of performance at the group level.

All regression design matrices included an intercept (choice bias or average RT), an autoregressive component (previous trial's choice or RT), and the transformed target and distractor coherence (scaled between -1 and 1). We included autoregressive components to capture well-established behavioral features like choice repetition and RT autocorrelation (Egner, 2007; Laming, 1979; Lau and Glimcher, 2005; Urai et al., 2019). We transformed feature coherences using a saturating nonlinearity,

$$coh^*_{feature} = \frac{tanh(\alpha_{feature} \times coh_{feature})}{tanh(\alpha_{feature})}$$

with α_{target} and $\alpha_{distractor}$ fit as free parameters. This nonlinearity was inspired by classical work on psychophysical scaling laws (i.e., Fechner–Weber–Stevens scaling, (Krueger, 1989; Nieder and Miller, 2003)), and more recent work demonstrating this scaling during cognitive control experiments (Servant et al., 2014; Stafford et al., 2011). This approach distinguishes the coherence nonlinearity (α) from how strongly coherence influences performance (β_{Choice} and β_{RT}), with our analyses focused on the latter. To constrain these α parameters, we estimated one parameter for both choice and RT, capturing similar nonlinearities across both performance measures.

In our more complex models (e.g., incentives), our primary focus was on how additional task features moderated the influence of tanh-transformed feature coherence on performance. Lower

order effects of moderating factors (e.g., previous distractor congruence) were included in the lapse rate for choice analysis, and as a main effect in RT analyses. The full parameter sets for all analyses are available in Supplementary Data.

We excluded trials in our regression if they were 1) the first trial of the block, 2) shorter than 200ms or longer than 2s, 3) occurred after an error or after a trial was too fast/slow and 4) in reaction time analyses, if the current trial was an error. These exclusion criteria were chosen to be inclusive, while avoiding trials where there were likely to be a mixture of different cognitive processes (e.g., post-error adjustments).

We performed statistical inference on the parameters using an estimate of the group-level error variance from the emfit package, necessary to avoid violations of independence across participants from our hierarchical modelling. Contrast tests across models used Welsh's (unequal variance) t-tests, with contrasts weighting studies by the square root of the sample size. We aggregated *p*-values across studies using Lipták's method (Lipták, 1958; Zaykin, 2011), weighting studies by the square root of their sample size. Correlations between parameters were calculated by converting the group-level MAP covariance matrix to a correlation matrix.

We generated posterior predictive checks (trend lines on figures) by generating regression model predictions for all trials, and then aggregating these predictions in the same way as participants' raw behavior. This approach allows us to distinguish whether our model systematically deviates from behavior from whether deviations are driven by variability in parameters across participants. To provide finer-grained insights into our model fit, we generated additional

posterior predictive checks that aggregate trends across all participants (Supplementary Figure 6) and that highlight single participants (Supplementary Figure 7). To provide further validation of the robustness of our parameter estimation procedure, we performed parameter recovery (simulated behavior from our best-fitting regression parameters, refit our model to this simulated behavior, and the compared data-generating and recovered parameters; Supplementary Figure 8) and parameter knock-out analyses (re-fit models with key nuisance regressors removed; Supplementary Figure 9). These robustness checks provided convergent evidence that our key parameters had good identifiability.

We generated sensitivity dynamics plots (e.g., Figure 6) by computing the regression-estimated coherence effect conditioned on RT. For a range of simulated RTs, the estimated motion sensitivity timeseries is:

$$\beta_{motion}^{RT} = (\beta_{motion} + \beta_{motion:RT} SimRT) \odot (1 - lapse^{RT})$$

$$lapse^{RT} = \frac{1}{1 + exp(-(\beta_{Lapse} + \beta_{\mathrm{RT}} SimRT))}$$

Where βs are regression weights estimated in our analysis, $SimRT$ is a vector of simulated RTs (e.g., .5:.01:1), and $\odot$ indicates element-wise multiplication. For control-dependent dynamics (i.e., incentivized dynamics; see Figure 7), we included 2-way and 3-way interactions between feature sensitivity, RT, and control drivers. We generated these sensitivity dynamics for each participant, and then plotted the mean and between-participant standard error.

Feedforward Inhibition with Control Model

To provide a bridge between our regression analyses and processes models of decision-making, we adopted a generative modeling approach and tested whether participant behavior could be reproduced by a sequential sampling model (Figure 8). This model was inspired by two theoretical traditions. The first was a classic connectionist model of cognitive control (Cohen et al., 1990), which demonstrated how top-down adjustments to target and distractor sensitivity in evidence accumulation framework can capture a wide range of behavioral phenomena. To capture apparent within-trial adjustments to feature processing (see Results), our second inspiration was from dynamical models of task set reconfiguration, both across-trial (Gilbert and Shallice, 2002; Musslick et al., 2018; Steyvers et al., 2019) and within-trial (Mante et al., 2013; Pagan et al., 2022). In these dynamic models, adjustments in feature gain behave as a dynamical system, starting at some initial condition and exponentially approaching a fixed point.

This model takes as inputs the color and motion coherence in support of different responses (e.g., $coh_{colorLeft}$), nonlinearly transforms these inputs (e.g., $coh^*_{colorLeft}$; see regression analyses above), and then integrates evidence for each response in separate rectified accumulators (x_{left} and x_{right}).

For example, evidence for the left response would be calculated as:

$$dx_{left} = -\lambda x dt + (\beta_{color} coh^*_{colorLeft} dt + \beta_{motion} coh^*_{motionLeft} dt + N(0, \sigma_x)_{left}\sqrt{dt}) - \omega(\beta_{color} coh^*_{colorRight} dt + \beta_{motion} coh^*_{motionRight} dt + N(0, \sigma_x)_{right}\sqrt{dt})$$

$$if\ x_{left} < 0;\ x_{left} = 0$$

The model makes a choice when one of the accumulators reaches a linearly collapsing decision bound rectified above 0.01. We used a balanced feedforward inhibition model without leak ($\lambda = 0$ and $\omega = 1$), approximating a (rectified) drift diffusion process (Bogacz et al., 2006). Note that parameterizations of a leaky competing accumulator could also approximate the DDM (Bogacz et al., 2007, 2006), and so are plausible alternatives to our implementation. We preferred the FFI model because it provides a simple interpolation between DDM and race-like decision processes.

To capture dynamics in participants' feature sensitivity, we modified our accumulation model to incorporate an attractor network for the feature weights (Mante et al., 2013), a model we call the feedforward inhibition with control model (FFIc model). In this model, control acts like a stochastic dynamical system. The system starts at an initial level of feature gain (β^0; e.g., due to bottom-up salience or learning). This feature gain exponentially approaches an asymptotic gain level (its 'fixed-point'; e.g., a setpoint on zero distractors gain), according to a decay rate K (e.g., control gain). For example, the motion gain would be governed by:

$$d\beta_{motion} = -\gamma\beta dt + K_{motion}(fixedpoint_{motion} - \beta_{motion})dt + N(0, \sigma_{gain})\sqrt{dt}$$

With the leak term γ fixed to 0 as in the decision process.

We simulated 10,000 trials for each combination of target discriminability and distractor congruence (11 x 11 x 10,000), and then aggregated simulated behavior in the same way we

aggregated participants' behavior. Simulation code and parameter sets are available at github.com/shenhavlab/PACT-public.

Transparency and openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study, and we follow APA Journal Article Reporting Standards (Appelbaum et al., 2018). All data and analysis code are available at github.com/shenhavlab/PACT-public. This study's design and its analysis were not pre-registered.

Results

Participants performed the Parametric Attentional Control Task (PACT), a perceptual discrimination task that required them to classify the dominant color in an array of moving dots (Figure 1a). Participants made bimanual responses, for example responding with their left hand when the dominant color was purple or blue or responding with their right hand when the dominant color was green or beige. To avoid stimulus repetition priming (Braem et al., 2019; Mayr et al., 2003), two colors were assigned to each response and the majority color did not repeat across sequential trials. Across trials, we varied the extent to which those dots were coherently moving in the same or opposite direction as the correct response (distractor interference; Experiments 1-3) and how easily the participant could determine the dominant color (target discriminability; Experiments 2-3; Figure 1b). Participants performed the main *Attend-Color* PACT in blocks of 100 trials. To enhance the potency of motion as a distracting dimension (Shiffrin and Schneider, 1977) and allow for additional measures of automaticity and feature-specificity, participants alternated between these blocks-of-interest and shorter blocks

(20-50 trials) in which participants instead responded to the direction of dot motion (*Attend-Motion* PACT; Figure 1c).

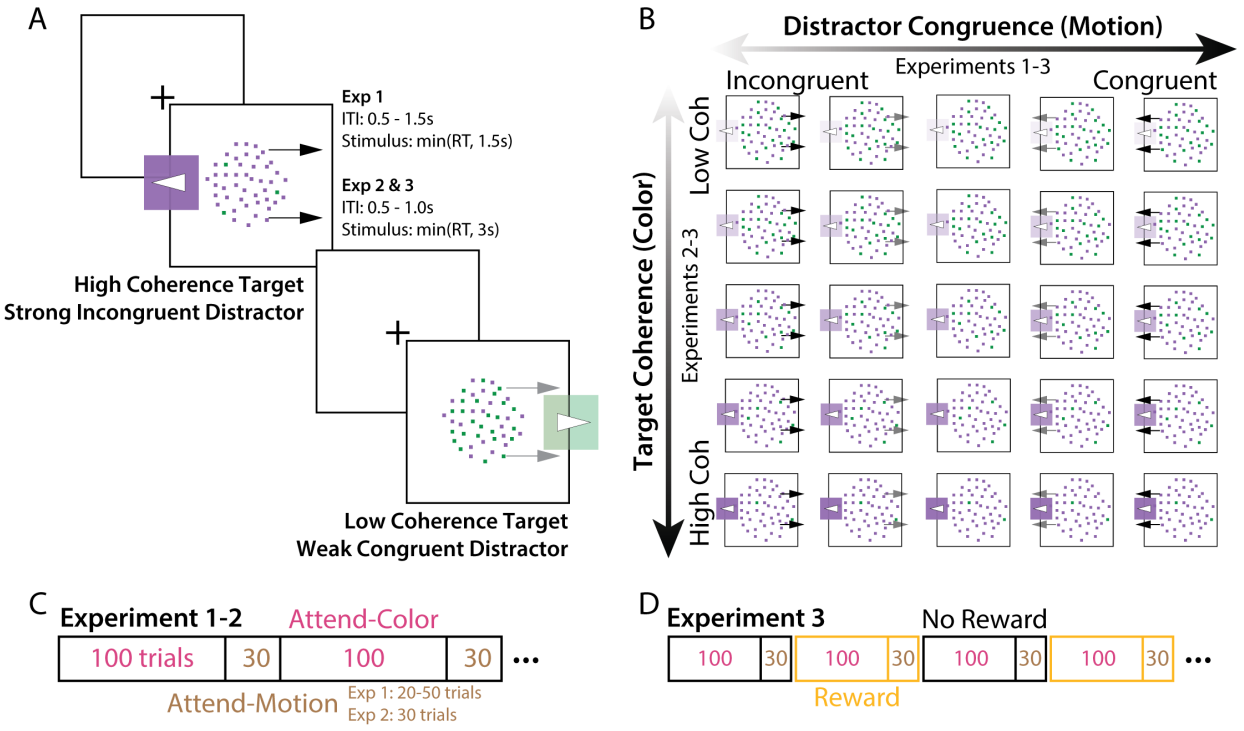


Figure 1. *Parametric Attentional Control Task (PACT).* **A)** On each trial, participants responded to the dominant color in a bivalent random dot kinematogram. This stimulus had a random color (target) coherence, depending on the proportion of dots that were in the majority. This stimulus also had a random motion (distractor) congruence, depending on motion coherence in the same or opposite direction as the color response. **B)** Across trials, we parametrically and independently varied the coherence of the dominant color (y-axis) and the congruence of the motion direction (x-axis). **C)** In Experiments 1 and 2, participants alternated between longer blocks of Attend-Color trials (target dimension was color, as in A) and shorter blocks of Attend-Motion trials (target dimension was motion). Participants took a self-timed break between blocks. **D)** In Experiment 3, participants alternated between pairs of Reward blocks and No Reward blocks. On Reward blocks, participants could earn a monetary bonus if they were fast and accurate, whereas we just encourage good performance on No Reward blocks. Participants were informed of the reward condition during their break between blocks.

Task performance varies parametrically with target discriminability and distractor interference

In Experiment 1 (N = 56), participants performed the PACT with uniformly colored dots (e.g., all blue or all green), but with the dots moving in a direction either congruent or incongruent with that target response. We varied the strength of this distractor dimension between being fully congruent with the correct color response (100% leftward coherence for a left color response) to being fully incongruent (100% rightward coherence for a left color response; Figure 1b). For trials mid-way between these two extremes (cf. 'neutral' trials), the dots did not move consistently in one direction or another (0% motion coherence).

Consistent with past research on cognitive control, we found that participants were slowest and least accurate when distractors were fully incongruent (median RT = 585ms, mean accuracy = 89%) and fastest and most accurate were fully congruent (median RT = 553ms, mean accuracy = 97%; cf. (Danielmeier et al., 2011). Performance on neutral trials (0% motion coherence) fell between these two extremes (median RT = 576ms, mean accuracy = 94%). Extending this work, hierarchical regression analyses (see Methods) revealed that performance varied in a graded fashion across this continuum of interference. Both accuracy (Cohen's *d* on regression estimate; $d = -1.47$) and reaction time ($d = 1.25$) worsened with parametrically increasing levels of interference ($ps < 0.001$, Figure 2c, Table 1).

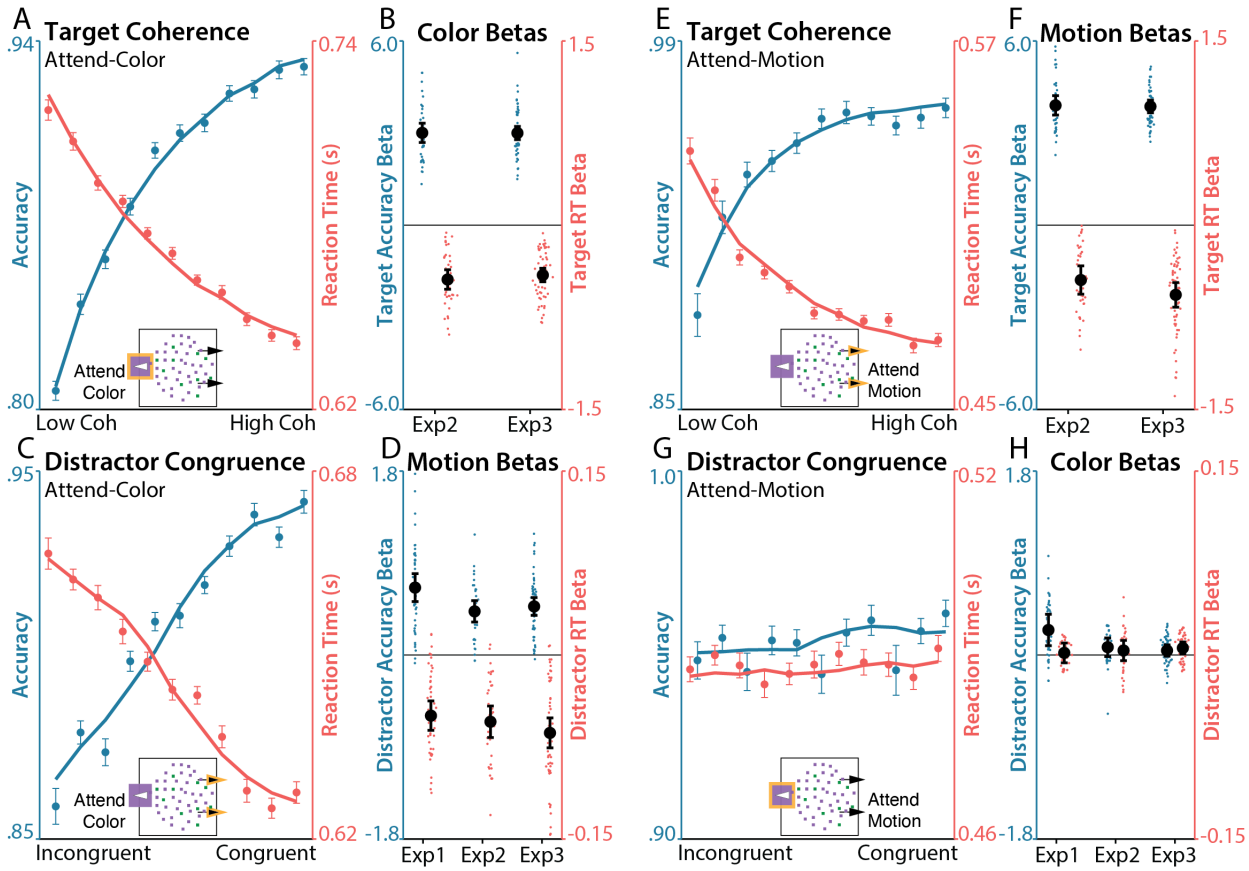


Figure 2. *Target and distractor sensitivity.* **A)** Participants were more accurate (blue, left axis) and responded faster (red, right axis) when the target color had higher coherence. Circles depict participant behavior and lines depict aggregated regression predictions. In all graphs, behavior and regression predictions are averaged over participants and experiments. Target sensitivity aggregated across Experiments 2 & 3. **B)** Regression estimates for the effect of target coherence on performance within each experiment, plotted for accuracy (blue, left axis) and RT (red, right axis). **C)** Participants were more accurate and responded faster when the distracting motion had higher congruence (coherence signed relative to target response). In all graphs, behavior and regression predictions are averaged over participants and experiments. Distractor sensitivity aggregated across Experiments 1-3. **D)** Regression estimates for the effect of distractor congruence on performance within each experiment, plotted for accuracy and RT. **E-F)** Similar to A-B, performance (E) and regression estimates (F) for the effects of target coherence during Attend-Motion blocks, in which motion was the target dimension. **G-H)** Similar to A-B, performance (G) and regression estimates (H) for the effects of distractor congruence during Attend-Motion blocks, in which color was the distractor dimension. Y-axis range is matched within-feature across tasks, see Supplementary Figure 10 for matched y-axes

across all features and tasks. Error bars on behavior reflect within-participant SEM, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization.

In Experiment 2 (N = 40) and Experiment 3 (N = 60), participants performed the same task as in Experiment 1, but we additionally varied the discriminability of the target (color) dimension. Across trials, the proportion of the majority color (*color coherence*) varied parametrically to make color discrimination easier (higher coherence) or more difficult (lower coherence). As in Experiment 1, the level of motion interference also varied across trials, independently of targets.

Consistent with past research on perceptual decision-making (Britten et al., 1992; Mante et al., 2013), we found that discrimination performance improved with higher levels of target discriminability. Participants in both studies were faster (Exp 2: d = -1.90, Exp 3: d = -1.99) and more accurate (Exp 2: d = 3.27, Exp 3: d = 3.73) with parametrically increasing levels of color coherence (aggregate ps < 0.001; Figure 2a, Table 1). At the same time, we continued to find that participants were slower and less accurate when the goal-irrelevant movement of those dots was increasingly incongruent with the correct color response (see Figure 2a, Table 1).

Performance on our task varied parametrically with both color coherence and motion coherence, but these two coherence manipulations were designed to exert their influence on performance in different ways. Whereas variability in color coherence was intended to influence the stimulus uncertainty directly relevant to goal-directed decision-making (i.e., determining which response is the correct one), motion coherence was intended to exert a more automatic influence on response selection by facilitating responses consistent with the direction of motion. We confirmed this assumption regarding the relative automaticity of motion versus color processing

by having participants perform interleaved blocks in which they responded based on motion and ignored color ('Attend-Motion'). We found that participants were more sensitive to the now-relevant motion coherence (Figure 2e), but were no longer sensitive to the now-irrelevant color congruence (Figure 2g; Supplementary Table 1-2). This asymmetry suggests that participants' decisions were not solely driven by the bottom-up salience of these features, as participants were more sensitive to color when it was relevant and less sensitive to motion when it was irrelevant, reflecting differential engagement of top-down control across the two tasks (Cohen et al., 1992).

Table 1. Target and distractor sensitivity

DV	**Predictors**	**Exp 1 (df = 45) Effect size (*d*)**	**Exp 2 (df = 25) Effect size (*d*)**	**Exp 3 (df = 45) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Target coherence		3.27	3.73	**1.01×10^{-44}**
	Distractor congruence	1.47	1.42	1.50	**4.89×10^{-32}**
	Target × Distractor		-0.184	-0.344	**0.0226**
RT	Target coherence		-1.90	-1.99	**1.59×10^{-28}**
	Distractor congruence	-1.25	-1.49	-1.43	**1.26×10^{-29}**
	Target × Distractor		0.230	0.0525	0.437

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

Target discrimination and distractor interference occur in parallel

We found that participants' task performance varied parametrically with both the target discriminability and distractor congruence, both for choice and reaction time. We next sought to further understand the relationships between these changes in performance, within and across features.

First, we tested whether a given feature exerted a similar influence on both accuracy and RT. We found that this was indeed the case, as there was a significant correlation between the effect distractors had on accuracy and RT (rs < -0.87, ps < 0.001). The influences of target discriminability on accuracy and RT were also significantly correlated (rs < -0.54, ps < 0.001; Supplementary Table 3). Thus, participants who became faster with higher levels of a given feature's strength also became more accurate, suggesting that accuracy and RT shared a common underlying process (e.g., evidence accumulation rate, which we return to below).

Second, we tested whether the influences of target discriminability and distractor congruence on performance were independent (e.g., distractors and targets are processed in parallel; (Lindsay and Jacoby, 1994; Servant et al., 2014) or instead modulatory (e.g., distractor congruence influences target sensitivity). If the two forms of feature processing modulated one another, we would predict that target and distractor coherence would interact in predicting performance. We did not find such an interaction in RTs (ds = 0.05 to 0.23, p = 0.33; Table 1), though we did find a small but significant interaction between target and distractor coherence on accuracy (ds = -0.18 to -0.34, p = 0.023). For both studies, removing target-distractor interactions as predictors in our accuracy regressions improved model fit (Protected exceedance probability on AIC: Exp 2 PXP = 1; Exp 3 PXP = 1). If distractors had an antagonistic influence on target processing, we would also predict that target and distractor sensitivity would be negatively correlated across subjects. Contrary to this prediction, these effects were either not significantly correlated or positively correlated, both for RT (Exp 2: $r(25) = .14$, $p = .48$; Exp 3: $r(45) = .44$, $p = .0019$) and accuracy (Exp 2: $r(25) = -.15$, $p = .45$; Exp 3: $r(45) = .12$, $p = .43$), suggesting that individual differences in target and distractor processing were not antagonistic.

Previous conflict preferentially suppresses distractor sensitivity

Within a given trial, we found that performance varies parametrically and independently with the coherence of target (color) and distractor (motion) features. We next sought to understand how participants adapted their information processing *across* trials, to provide insight into the control processes that guide performance in this task. We measured how participants' feature sensitivity changed after difficult (e.g., more incongruent) trials, an index of cognitive control known as conflict adaptation (Egner, 2007; Gratton et al., 1992). The classic effect is that participants show weaker congruence effects after incongruent trials than after congruent trials, with the traditional interpretation being that this reflects upregulated target sensitivity (Botvinick et al., 2001; Egner, 2007). Our task allowed us to build on this work to test whether this adaptation effect varies parametrically with distractor congruence. Critically, we can also test whether adaptation occurs through an influence of previous conflict on subsequent target enhancement, distractor suppression, or both. Finally, we can further test whether adaptation occurs due to the discriminability of the *target* on the previous trial.

Across all three of our studies, we found that participants' sensitivity to the distractor dimension was robustly and parametrically influenced by the distractor congruence on the previous trial, as reflected both in their choice (ds = 1.44 to 1.74, $p < .001$; Figure 3a) and RT (ds = 0.83 to 1.79, $p < .001$; Figure 3b; Table 2). When the previous trial had congruent distractors, participants had strong sensitivity to the distractor congruence (Figure 3a-b, navy). When the previous trial had incongruent distractors, participants were much less sensitive to distractors (Figure 3a-b, red). These patterns are consistent with those typically observed in studies of conflict adaptation

(Danielmeier et al., 2011; Egner, 2007), and further demonstrate gradations within these classic effects.

When varying both target and distractor features (Experiments 2-3), we found an additional influence of previous distractor congruence on target processing, whereby more incongruent previous trials enhanced the influence of target discriminability on the current trial (Figure 3d-e). However, the influence of previous distraction on target processing was substantially smaller than its effect on distractor processing (see Figure 5), and was only found for accuracy ($p < .001$) and not RT ($p = .57$), Finally, we found that performance adapted to the strength of the previous *target*, with less-discriminable targets yielding *lower* sensitivity to target strength (i.e., poorer performance) on the following trial, potentially due to disengagement (Supplementary Figure 2, Table 2). However, like the distractor-target effect, this target-target effect was much smaller than the distractor-distractor effects and only observable in accuracy ($p < .001$) and not RT ($p = .19$).

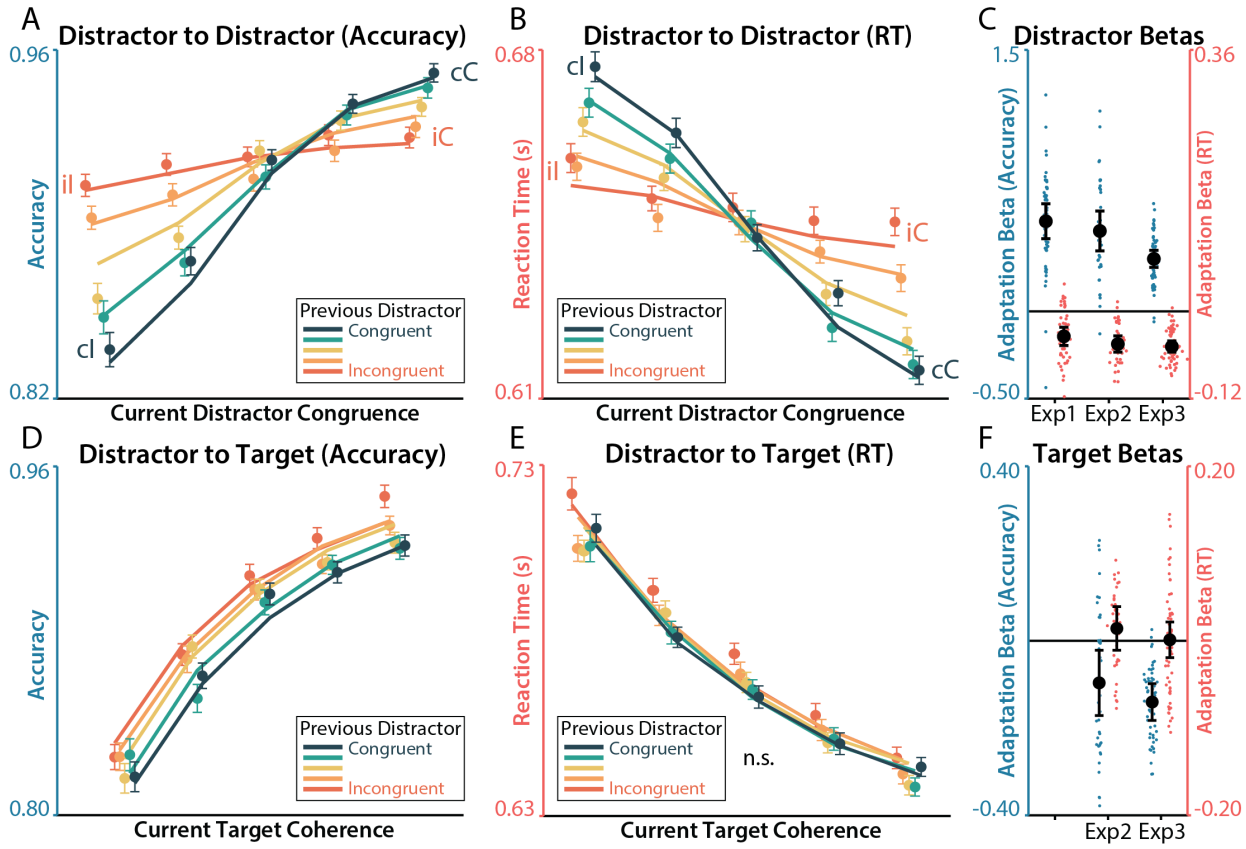


Figure 3. *Distractor-dependent adaptation.* **A-B)** The relationship between distractor congruence and accuracy (A) and RT (B) was weaker when the previous trial was more incongruent (redder colors). Circles depict participant behavior and lines depict aggregated regression predictions. **C)** Regression estimates for the current distractor congruence by previous distractor congruence interaction, within each experiment. **D-E)** The relationship between target coherence and performance was stronger after more incongruent trials in accuracy (D) but not RT (E). **F)** Regression estimates for the current target coherence by previous distractor congruence interaction, within each experiment. Error bars on behavior reflect within-participant SEM, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization. Feature coherence was rank-ordered and binned into quantiles with equal numbers of trials at each level of target coherence, distractor congruence, or previous distractor congruence.

Table 2. Effects of previous conflict on feature sensitivity					
DV	**Predictors**	**Exp 1 (df = 41) Effect size (*d*)**	**Exp 2 (df = 15) Effect size (*d*)**	**Exp 3 (df = 35) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Distractor × Prev Distract	1.59	1.45	1.74	**6.15×10^{-31}**
	Distractor × Prev Target		-0.670	0.103	0.964
	Target × Prev Distract		-0.473	-0.990	**2.83×10^{-8}**
	Target × Prev Target		0.418	0.644	**1.25×10^{-5}**
Lapse Rate	Prev Distract	-0.522	-0.498	-1.04	**1.75×10^{-10}**
	Prev Target		-0.110	-0.494	**0.00934**
RT	Distractor × Prev Distract	-0.836	-1.44	-1.79	**8.99×10^{-24}**
	Distractor × Prev Target		0.174	0.0618	0.520
	Target × Prev Distract		0.210	0.0155	0.726
	Target × Prev Target		0.147	0.154	0.285
	Prev Distract	0.287	0.202	-0.267	0.623
	Prev Target		0.109	-0.275	0.0884

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

A common concern when measuring conflict adaptation effects is the extent to which these reflect control adjustment (as typically assumed) or low-level priming that can occur due to stimulus-stimulus or stimulus-response associations (Braem et al., 2019; Hommel et al., 2004; Mayr et al., 2003; Schmidt and De Houwer, 2011). For example, in some tasks, if two adjacent trials are both congruent or both incongruent, they are also more likely to share stimulus-response mappings, biasing analyses of sequential adaptation (Schmidt, 2019). Our experiment was designed to largely avoid potential priming confounds by eliminating stimulus repetitions (with two colors assigned to each response hand that never repeat), and by using stochastic

motion stimuli (versus, e.g., static arrows) that also have very infrequent exact repetitions. For example, the probability that two trials will have the same motion coherence was only 9%.

However, to further rule out that our key adaptation findings resulted from priming effects, we tested whether adaptation effects were present in our more automatic Attend-Motion blocks. Whereas a priming account would predict similar (within-feature) adaptation effects across both Attend-Color and Attend-Motion blocks (Moeller and Frings, 2014), a cognitive control account would predict weaker adaptation effects for Attend-Motion than Attend-Color blocks. We found that adaptation effects during Attend-Motion blocks were overall weak and inconsistently signed (e.g., previous interference led to either increased or decreased sensitivity to distractors across studies; Supplementary Table 4-5). Comparing the adaptation effects across the two types of blocks directly, we found significantly stronger adaptation effects during Attend-Color than Attend-Motion blocks. Distractor adaptation was weaker during Attend-Motion than Attend-Color, despite including color repetitions during Attend-Motion blocks (Choice: $p < .001$; RT: $p < .001$). Critically, we can directly compare trial-to-trial changes in motion sensitivity when motion is task-relevant (Attend-Motion) and task-irrelevant (Attend-Color), matching the salience of this motion dimension across tasks (Giesen et al., 2012). Target adaptation during Attend-Motion was not significant (Choice: $p = .268$; RT: $p = .777$; Supplementary Table 4) and was weaker than distractor adaptation during Attend-Color (Choice: $p < .001$; RT: $p = .34$; Supplementary Table 5). Together, these results suggest that the adaptation effects we observed during Attend-Color trials likely reflected changes in control states rather than stimulus-driven priming.

In addition to influencing sensitivity of choices and RTs to individual features (adaptation effects described above), we found that previous target and distractor information also exerted a small but reliable influence on the likelihood that the participant would respond randomly on the next trial (*lapse rate,* see the Regression Analysis subsection in Methods). Specifically, higher levels of distractor incongruence and lower levels of target discriminability increased subsequent lapse rates (*p*s < .001; Table 2), though these changes were subtle (e.g., post-congruent lapse rates ranged from 0.023% to 0.13% across studies; post-incongruent lapse rates ranged from 0.13% to 0.41% across studies). We did not otherwise find consistent main effects of previous targets and distractors on choice behavior (i.e., in the direction of a particular response) or on RT.

Performance incentives preferentially enhance target sensitivity

We found that performance on our task adapted to previous distractor-related interference, and that this influence was observed primarily in subsequent processing of the (motion) distractor rather than the (color) target. This may reflect a fundamental bias in the control system towards adjusting distractor processing in our task, but it may also reflect a process that is specialized for conflict adaptation. To disentangle these possibilities, we examined how target and distractor processing are influenced by heightened levels of motivation. In Experiment 3 we incorporated an incentive manipulation, with blocks of trials for which participants could either earn a monetary reward for fast and accurate performance, and blocks where performance was not rewarded (Figure 1d).

We found that participants' accuracy was more sensitive to target discriminability in rewarded blocks than non-rewarded blocks ($d = 0.61$, $p < .001$; Figure 4a, Table 3). This effect of

incentives on target sensitivity was specific to choice and not RTs (d = -0.10, p = 0.47), though participants were overall faster in rewarded blocks (d = -0.41, p = 0.0045). Participants were also marginally more likely to make lapses responses during rewarded blocks (d = 0.244, p = 0.092). In terms of distractors, we found that in rewarded blocks participants were less sensitive to distractors in RT (d = 0.35, p = 0.012), albeit with a small effect size, and that incentives did not significantly influence distractor sensitivity in choice (d = -0.016, p = 0.91).

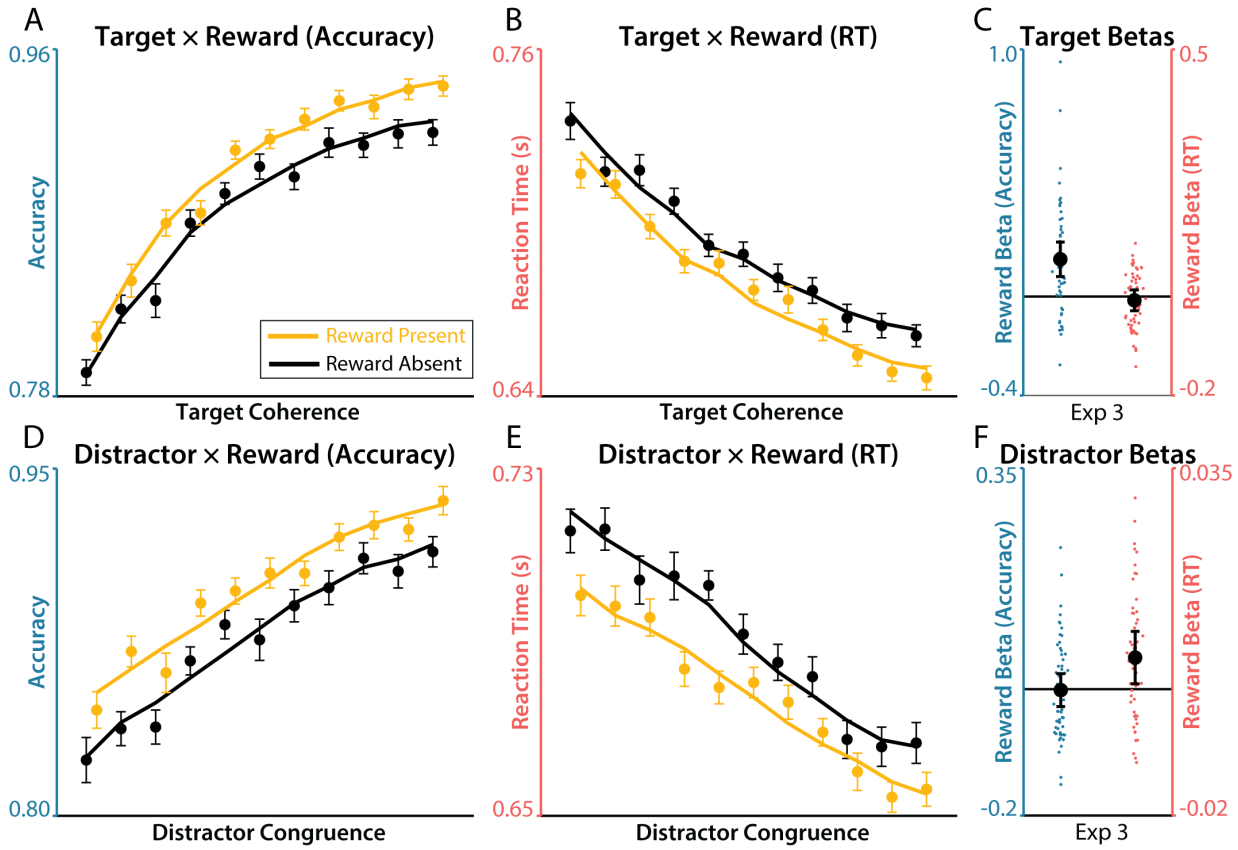


Figure 4. *Influence of incentives on target and distractor sensitivity.* **A-B)** The relationship between target coherence and performance was stronger during incentivized blocks (gold) in the domain of accuracy (A), but not RT (B). Circles depict participant behavior and lines depict aggregated regression predictions. **C)** Regression estimates for the target coherence by incentive interaction. **D-E)** The relationship between distractor congruence and performance was weaker on incentivized blocks (gold) in the for RT (E), but not Accuracy (D). **F)** Regression

estimates for the distractor congruence by incentive interaction. Error bars on behavior reflect within-participant SEM, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization.

We further found that the target-enhancing effects of incentives also were not specific to the color dimension. When motion was the target dimension (attend-motion blocks), incentives preferentially increased sensitivity to motion coherence ($d = 0.70$, $p < .001$). Interestingly, incentives had an even larger influence on target sensitivity in attend-motion relative to attend-color blocks ($t(59.0) = 2.14$, $p = 0.036$; Supplementary Table 6-7).

Table 3. Effects of incentives on feature sensitivity

DV	**Predictors**	**Exp 3 (df = 41) Effect size (*d*)**	***p*-value**
Choice	Target × Reward	0.612	**8.56×10^{-5}**
	Distractor × Reward	-0.0156	0.911
Lapse Rate	Reward	0.244	0.0924
RT	Target × Reward	-0.103	0.467
	Distractor × Reward	0.349	**0.0195**
	Reward	-0.411	**0.00447**

Effect sizes are calculated from MAP group-level regression estimates. Statistically significant *p*-values (two-tailed, $\alpha = 0.05$) are shown in bold.

Previous conflict and incentives have dissociable influences on target and distractor processing

Our within-trial results demonstrated that participants are sensitive to target (color) and distractor (motion) information, with little interaction between these dimensions. Consistent with this

putative independence, we found that previous interference primarily influenced distractor sensitivity (suppressing distractor sensitivity after trials with incongruent distractors), and that rewards primarily influenced target sensitivity (enhancing target sensitivity when incentivized). These findings strongly suggest a dissociation between target and distractor processing.

To confirm these findings, we formally tested the double dissociation between how incentives and previous interference influenced target and distractor choice sensitivity (Figure 5). We found that previous conflict had a larger absolute effect on distractor processing than it did on target processing in both accuracy ($t(31.4) = 9.54$, $p = 8.36 \times 10^{-11}$) and RT ($t(33.7) = 4.64$, $p = 5.14 \times 10^{-5}$). We found that rewards conversely had a larger influence on targets than distractors in Accuracy ($t(44.5) = 5.08$, $p = 7.22 \times 10^{-6}$), though not in RT ($t(37.7) = 0.25$, $p = 0.80$). Critically, the difference-of-differences was also significant in both Accuracy ($t(39.6) = 10.2$, $p = 1.36 \times 10^{-12}$) and RT ($t(48.3) = 3.11$, $p = 0.0031$), supporting dissociable control over different dimensions of feature processing.

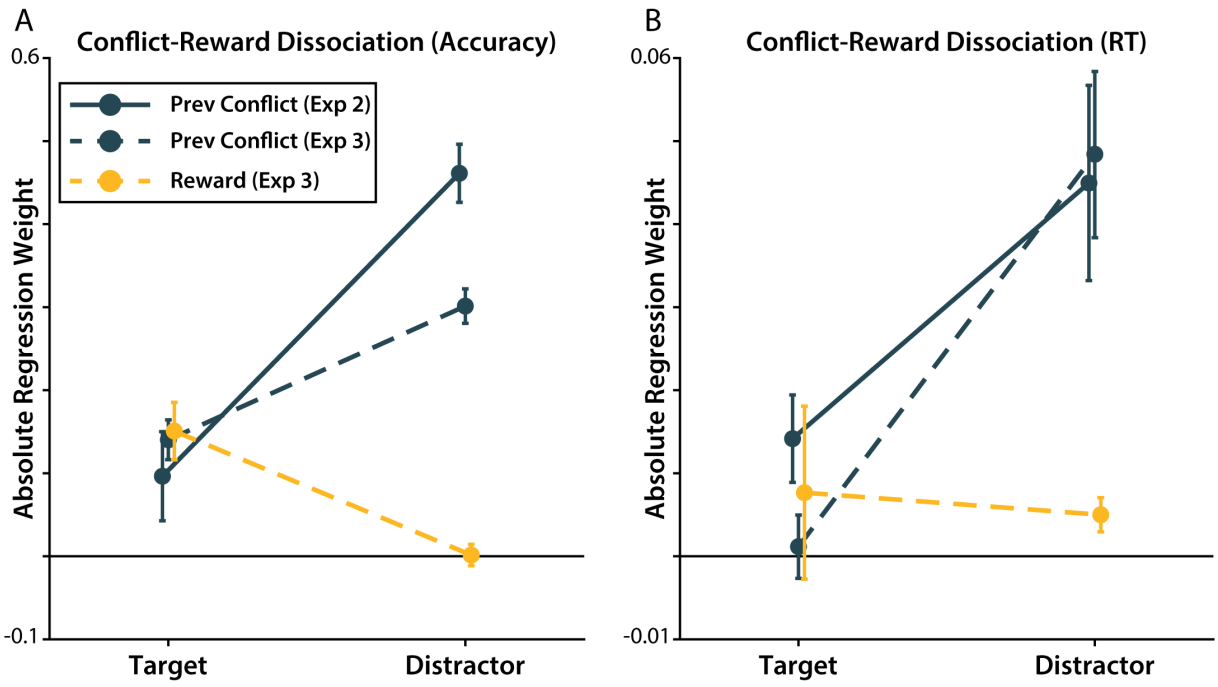


Figure 5. *Dissociations between previous conflict and incentive effects.* Post-conflict effects were significantly larger on distractor sensitivity than target sensitivity in accuracy (A) and RT (B). In contrast, reward effects were significantly larger on target sensitivity than distractor sensitivity in accuracy (A) and similarly large in RT (B). Errors bars show MAP SEM.

These findings are consistent with a previous neuroimaging experiment that found incentives enhanced responses in target-related areas (visual word form area for text targets) and mostly-incongruent blocks suppressed responses in distractor-related areas (fusiform face area for face distractors; (Soutschek et al., 2015). In the following sections, we extend these convergent findings to explore how previous conflict and incentives influence the dynamics of control implementation.

Differential within-trial dynamics of target and distractor processing

Our initial results show that participants independently control their sensitivity to target (color) and distractor (motion) information. However, previous research has revealed that participants' task processing also dynamically changes within a trial (Servant et al., 2014; Weichart et al., 2020; White et al., 2011), including in response to incentives (Adkins and Lee, 2021). Whereas much of the previous research has focused on dynamics in spatial attention during flanker tasks (e.g., a shrinking spotlight of attention; (Weichart et al., 2020; White et al., 2011), less is known about the dynamics of attention between features of conjunctive stimuli like those in our task, where target and distractor processing may be more independent (Adkins and Lee, 2021; Servant et al., 2014).

To test how sensitivity to target and distractor features changed within each trial, we measured whether the influence of coherence on participants' choices depended on reaction time (i.e., the choice ~ coherence × RT interaction). These analyses work under the logic that faster RTs reflect earlier epochs of information processing, which we confirm through subsequent evidence accumulation simulations (see 'An accumulator model of attentional control over target and distractor processing' in Results; Supplementary Figures 4-5). Our approach builds on 'delta function' analyses of how congruence effects differ across RT quantiles (De Jong et al., 1994; Ridderinkhof, 2002; van den Wildenberg et al., 2010)[3], extended this methodology with a GLM approach that estimates parametric changes in both target and distractor sensitivity over time.

[3] Previous work on delta-plot analyses have investigated how RT difference scores (e.g., congruent – incongruent) vary across RT quantiles. This work has been criticized based on the inherent mean-variance relationship in skewed RT distributions (Zhang and Kornblum, 1997). Instead, our analyses investigate how accuracy effects vary as a function of RT instead, avoiding this concern.

At the earliest RTs, participants were the least sensitive to targets (Figure 6a) and the most sensitive to distractors (Figure 6d). At later RTs, participants became more sensitive to targets (ds = 0.69 to 0.97, $p < .001$), and less sensitive to distractors (ds = -0.71 to -1.5, $p < .001$; Table 4). This is consistent with an attentional control process that enhances sensitivity to goal-relevant features and suppresses attention towards goal-irrelevant features. Notably, these results suggest that this attentional process occurs 'online' within the course of a trial.

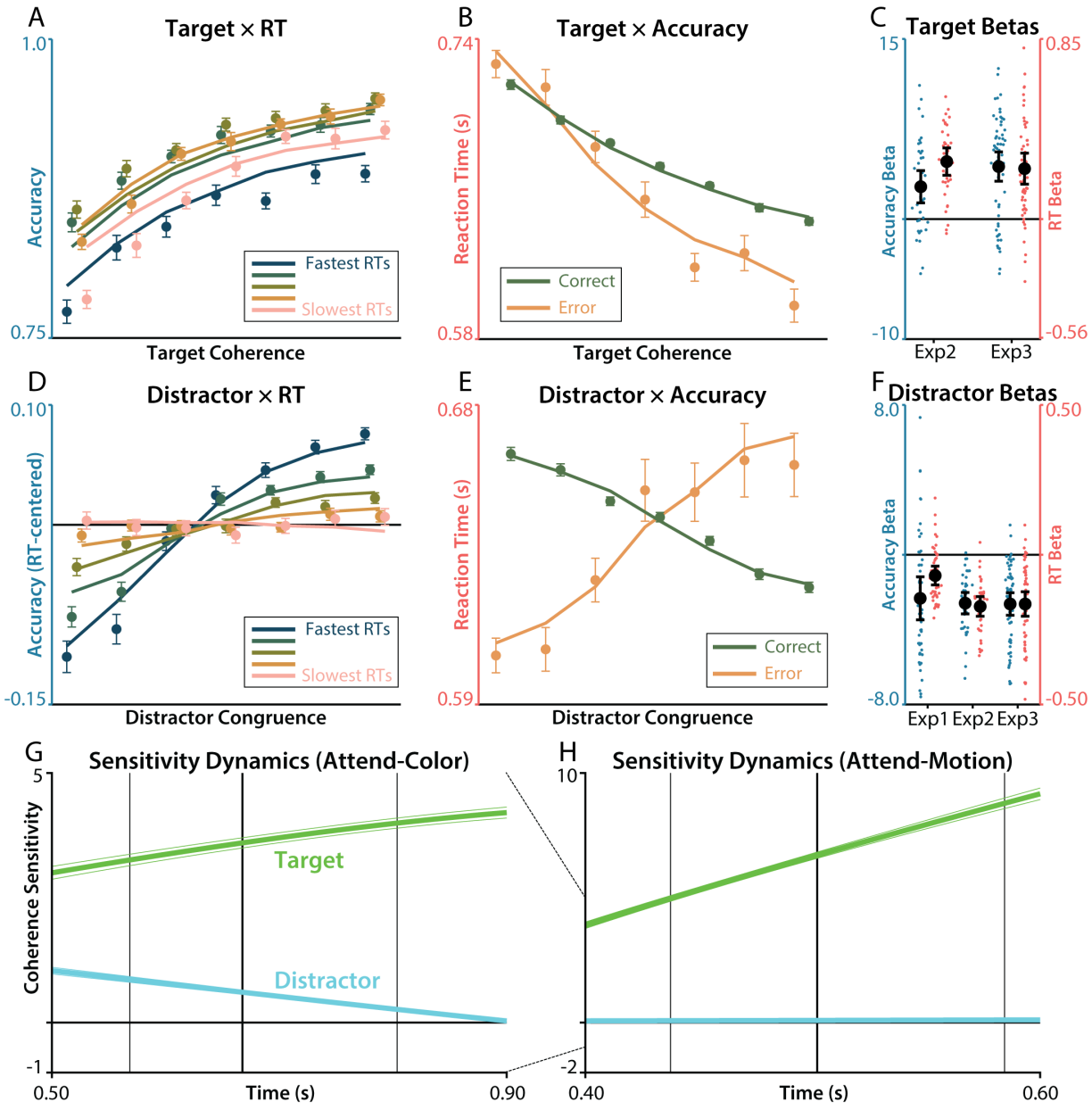


Figure 6. *Target and distractor sensitivity dynamics.* **A)** The relationship between target coherence and accuracy increased at later RTs (pinker color). **B)** Participants responded faster on error trials that correct trial when target coherence was higher. **C)** Regression estimates for the interaction between target coherence and RT (blue) and accuracy (red), within each experiment. **D)** The relationship between distractor congruence and accuracy decreased at later RTs (pinker). Note that these data are mean-centered within each RT bin to remove the target effects in (A) from this visualization of distractor sensitivity. **E)** Participants responded faster on error trials than correct trials when distractors were incongruent. **F)** Regression estimates for the interaction between distractor congruence and

RT (blue) and accuracy (red), within each experiment. **G)** Target (green) and distractor (cyan) sensitivity plotted as a function of reaction time, as estimated by our regression model in Attend-Color blocks. Vertical lines indicate quartiles of the RT distribution. **H)** Same as G, but generated from regression models fit to the Attend-Motion blocks. Note the different scaling of the x-axis and y-axis (see dashed line between plots). Error bars on behavior reflect within-participant SEM, error bars on sensitivity estimates reflect between-participant SEM of the predictions, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization. Feature coherence and RT were rank-ordered and binned into quantiles with equal numbers of trials at each level of target coherence, distractor congruence, or RT bin.

We also fit a complementary analysis for RT (i.e., the *RT ~ coherence × accuracy* interaction). We found that participants had steeper target coherence slopes on error trials (*ds* = 0.89 to 1.5, $p < .001$; Figure 6b), driven by faster errors when the targets were high coherence, consistent with participants responding before their maximal target sensitivity. Likewise, we found that the relationship between RT and distractor congruence inverted on error trials (*ds* = -0.68 to -1.8, $p < .001$; Figure 6e), with participants making faster errors on more incongruent trials, consistent with an early sensitivity to distractors that is suppressed over time.

Table 4. Dynamics of feature sensitivity across response times

DV	Predictors	Exp 1 (df = 38) Effect size (*d*)	Exp 2 (df = 21) Effect size (*d*)	Exp 3 (df = 41) Effect size (*d*)	Aggregate *p*-value
Choice	Target × RT		0.686	0.975	**4.61×10^{-11}**
	Distractor × RT	-0.709	-1.54	-1.19	**8.40×10^{-20}**
Lapse Rate	RT	0.247	0.928	0.534	**9.25×10^{-13}**
RT	Target × Accuracy		1.46	0.889	**1.99×10^{-14}**
	Distractor × Accuracy	-0.683	-1.82	-1.09	**1.39×10^{-20}**

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

These findings suggest online dynamics in the allocation of top-down attention to facilitate target processing and suppress distractor processing, but it is possible that they instead reflect dynamics inherent to the bottom-up processing of color and motion information. To rule out this alternative hypothesis, we tested whether similar sensitivity dynamics were present during Attend-Motion blocks, when color information serves as a much less potent distractor. During these blocks, we found that participants enhanced target (motion) sensitivity faster than they did during Attend-Color blocks ($p < .001$; Figure 6h; Supplementary Table 8-9). In contrast, participants had slower distractor sensitivity dynamics during Attend-Motion blocks ($p < .001$). Together these results demonstrate that these sensitivity dynamics depend on the task that participants are performing, rather than being exclusively due to stimulus-driven factors.

Finally, we tested whether participants' within-trial attentional dynamics changed over the course of the experiment, modeling the linear change in parameters across blocks of trials. We found that later in the experiment, participants' overall sensitivity to distractors was higher (in choice), and their sensitivity to targets was lower (in reaction time; see Supplementary Table 11). However, later in the experiment participants also had faster target and distractor dynamics, such that maladaptive sensitivity was most prominent in the earliest phase of the trial (Supplementary Figure 11). These results speculatively suggest that over time participants shift from maintaining initial sensitivity to reactively reconfiguring attention, potentially due to fatigue or proactive interference from Attend-Motion blocks.

Previous conflict and incentives influence early trial dynamics

We found that, within a trial, participants dynamically adjusted attention depending on the task at hand, with increasing sensitivity to task-relevant color information and decreasing sensitivity to task-irrelevant motion information over the course of a trial. This raises the question whether the two forms of adaptation we observed, related to previous conflict and incentives, influenced different components of the within-trial attentional dynamics.

To address this question, we first examined how the dynamics of target and distractor sensitivity were altered by the congruence of the distractor on the previous trial (i.e., *Choice* ~ *PreviousDistractor* × *RT* × *Coherence*). We found that after incongruent trials, participants started the next trial more sensitive to targets and less sensitive to distractors (Figure 7a). Although this means that after congruent trials participants had worse initial conditions (starting less sensitive to targets and more sensitive to distractors), they appeared to compensate for this early disadvantage with faster increases in target enhancement (ds = 0.65 to 1.0, $p < .001$) and distractor suppression (ds = -0.68 to -1.1, $p < .001$; Table 5). Both post-congruent and post-incongruent trials thus reached similar asymptotic levels of feature sensitivity. This early influence of previous conflict on congruence sensitivity is consistent with previous experiments on the timecourse of conflict adaptation (Stins et al., 2008; Wylie et al., 2010), with the current work extending these findings to show concurrent, albeit weaker, target-enhancement dynamics.

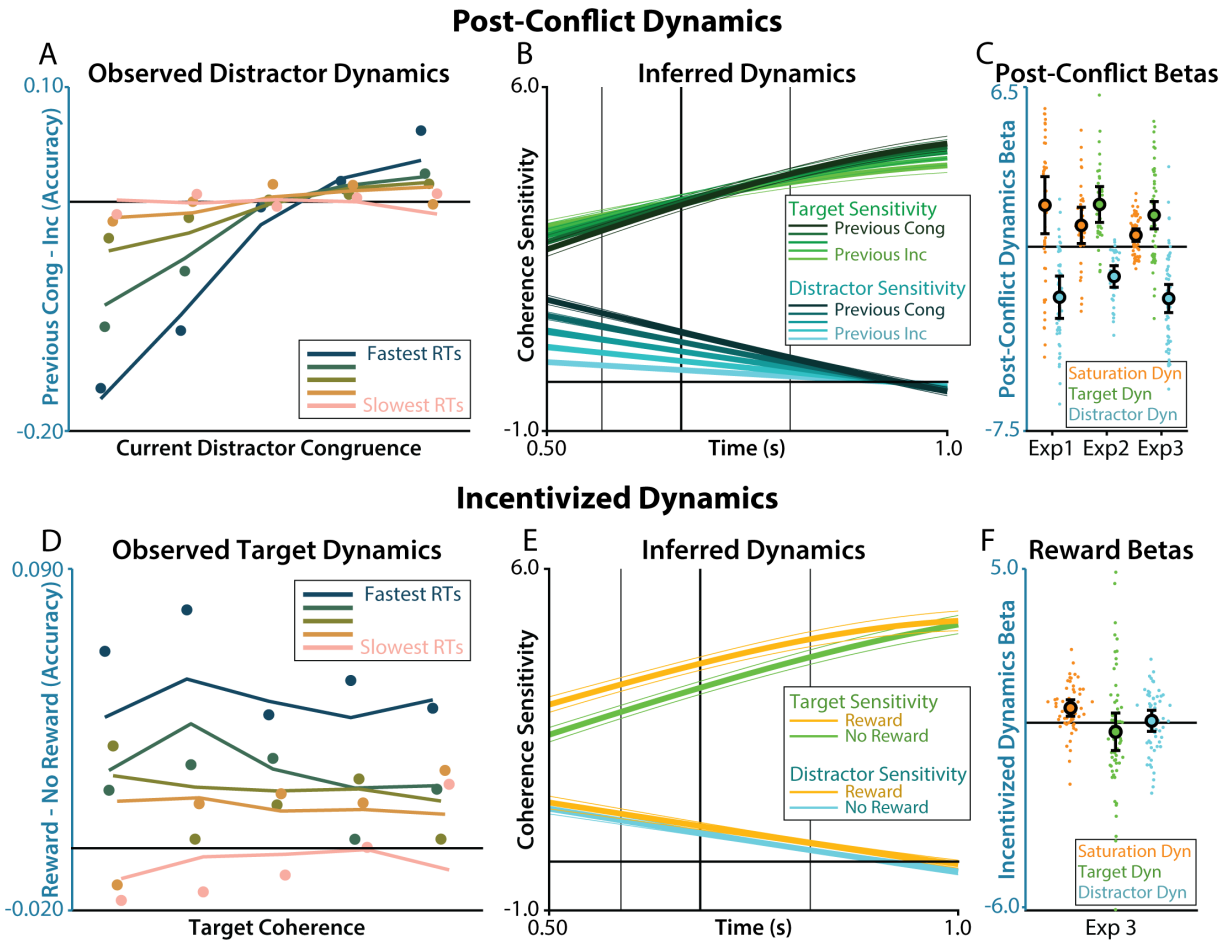


Figure 7. *Influence of conflict and incentives on sensitivity dynamics.* **A)** The relationship between previous distractor congruence and current distractor congruence was strongest for early RTs (bluer color). The y-axis depicts the difference in accuracy between the extreme tertiles of previous congruence, for visualization purposes. **B)** Target (green) and distractor (cyan) sensitivity plotted as a function of previous congruence (color shade) and reaction time (x-axis), as estimated by our regression model. Vertical lines indicate quartiles of the RT distribution. **C)** Regression estimates for the interactions between reaction time and previous congruence on lapse rate ('Saturation Dynamics', orange); or reaction time, previous congruence, and feature coherence on accuracy (target is green, distractor is cyan). **D)** The relationship between incentives and target coherence was strongest for early RTs (bluer color). The y-axis depicts the difference in accuracy between blocks where there were rewards vs blocks without rewards. **E)** Target (green) and distractor (cyan) sensitivity plotted as a function of incentives (gold) and reaction time (x-axis), as estimated by our regression model. Vertical lines indicate quartiles of the RT distribution. **F)** Regression estimates for the interactions between reaction time and incentives on lapse rate (orange); or reaction time,

incentives, and feature coherence on accuracy (target is green, distractor is cyan). Error bars on behavior reflect within-participant SEM, error bars on sensitivity estimates reflect between-participant SEM on the predictions, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization. Feature coherence and RT were rank-ordered and binned into quantiles with equal numbers of trials at each level of target coherence, distractor congruence, or RT bin.

Figure 5. Effects of previous conflict on feature sensitivity dynamics					
DV	**Predictors**	**Exp 1 (df = 26) Effect size (*d*)**	**Exp 2 (df = 9) Effect size (*d*)**	**Exp 3 (df = 29) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Prev Dist × Dist × RT	-0.853	-1.07	-1.01	**1.62×10^{-14}**
	Prev Dist × Targ × RT		1.01	0.646	**1.71×10^{-7}**
Lapse Rate	Prev Dist × RT	0.456	0.463	0.531	**3.52×10^{-6}**
RT	Prev Dist × Dist × Acc	-0.563	-1.15	-1.16	**2.84×10^{-13}**
	Prev Dist × Targ × Acc		-0.00500	-0.524	0.135
	Prev Dist × Acc	0.417	0.175	-0.162	0.881

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

We performed the equivalent analysis for incentive-related adaptation (i.e., *Choice ~ Reward × RT × Coherence*). We found that during incentivized blocks, participants' initial target sensitivity was higher than during non-incentivized blocks, and remained so across much of the trial (see Figure 7d). However, target sensitivity eventually reached an asymptote, such that towards the end of the trial both incentivized and non-incentivized trials had similar levels of target sensitivity (see slowest quantile in Figure 7d). This convergence was accounted for by larger increases in lapse rates later in incentivized trials ($d = 0.52$, $p < .001$; Table 6). The dynamics of distractor sensitivity, by contrast, did not significantly differ between incentivized and non-incentivized trials ($d = 0.055$, $p = 0.71$).

Table 6. Effects of incentives on feature sensitivity dynamics			
DV	**Predictors**	**Exp 3 (df = 29) Effect size (*d*)**	***p*-value**
Choice	Rew x Target × RT	-0.139	0.331
	Rew x Distractor × RT	0.0546	0.712
Lapse Rate	Rew × RT	0.524	**0.000937**
RT	Rew × Target × Acc	0.437	**0.00432**
	Rew × Distractor × Acc	-0.170	0.260
	Rew × Acc	-0.540	**0.000646**

Effect sizes are calculated from MAP group-level regression estimates. Statistically significant *p*-values (two-tailed, $\alpha = 0.05$) are shown in bold.

An accumulator model of attentional control over target and distractor processing

Our results demonstrate that participants independently control the initialization and online adjustment of attention towards target and distractor features. To parsimoniously account for this set of findings, we developed an accumulator model that integrated elements of previous models used to separately account for performance in tasks involving perceptual discrimination (Gold and Shadlen, 2007; Ratcliff and McKoon, 2008) and overriding prepotent distractors (Cohen et al., 1990; Weichart et al., 2020; White et al., 2011). We used a variant of a feedforward inhibition model, in which inputs provide excitatory inputs to associated response units and inhibitory inputs to alternative response units (Shadlen and Newsome, 2001). Our decision model takes as inputs the color and motion coherence in support of different responses, nonlinearly transforms these inputs, and then integrates evidence for each response in separate

rectified accumulators with balanced feedforward excitation and inhibition (Figure 8). The signal-to-noise ratio of the intermediate layer's outputs are determined by control units that determine the gain of a given feature (Cohen et al., 1990; Musslick et al., 2019). We hand-tuned the parameters of this model to determine whether it could capture our core experimental findings across choice and reaction time.

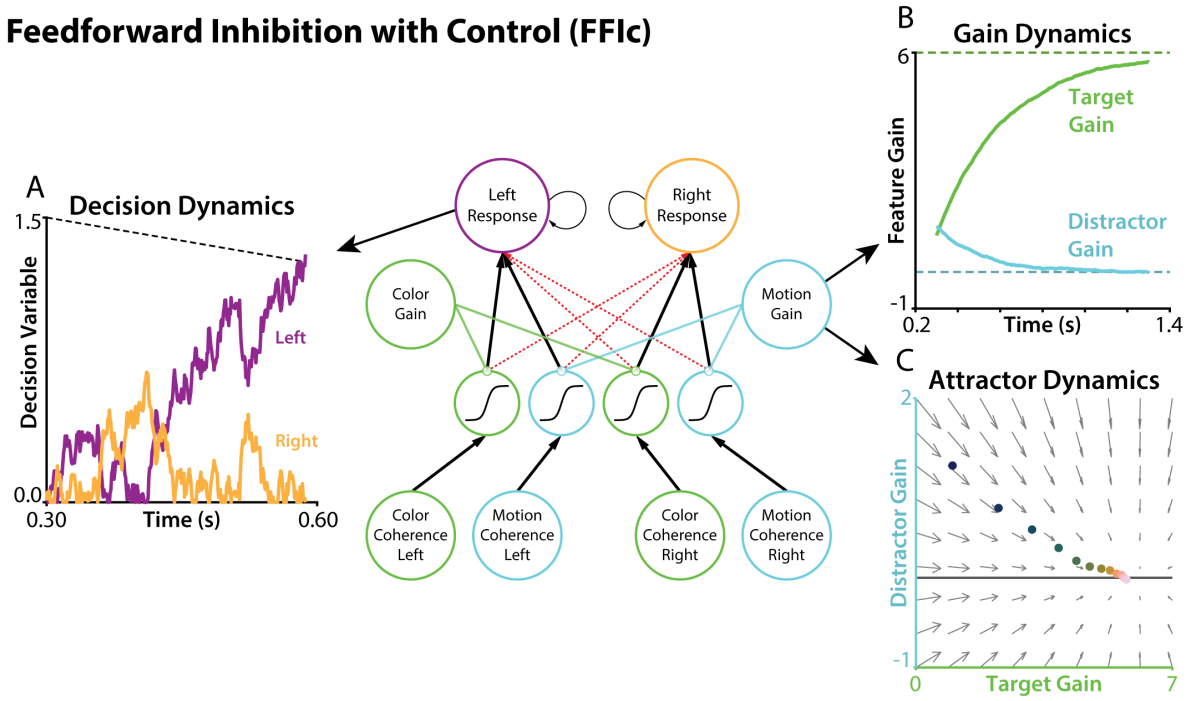


Figure 8. *Feedforward inhibition with control.* Color evidence (green) and motion evidence (blue) are transformed and accumulated to make a choice. Balanced excitatory connections (black solid lines) and inhibitory connections (red dashed lines) cause accumulation of the difference in evidence for each response. **A)** Evidence for the left response (purple) and right response (orange) are accumulated over time without leak. When one of the accumulators crosses a (linearly collapsing) decision threshold, the model chooses that response. **B)** Within each trial, the signal-to-noise of each feature pathways is controlled by a feature gain. Over time within a trial, the feature gains for targets (green) and distractors (cyan) exponentially approach to a fixed level (high gain for targets, zero gain for distractors). Note the difference in x-axis scaling compared to Figure 6G. **C)** An equivalent visualization of

the dynamics in B. Attractor dynamics drive target and distractor gains to their fixed level, shown at different timepoints within the trial (pinker colors are later in the trial). The horizontal line depicts zero distractor gain.

Our accumulator model was able to reproduce our key within-trial findings. During our main Attend-Color trials, it generated responses that were faster and more accurate with increasing color coherence (Figure 9a) and slower and less accurate with increasing motion incongruence (Figure 9b). We simulated Attend-Motion trials by increasing the target gain and decreasing the distractor gain, to capture potential differences in both automaticity and control. Now, our model generated responses that were even faster and more accurate with increasing target coherence (now motion; Figure 9c) but that were insensitive to distractor congruence (now color; Figure 9d), replicating the main behavioral results in Attend-Motion blocks. Notably, distractor effects were not reproduced in an accumulator competition model parameterized to be more 'race-like' (Supplementary Figure 3; (Teodorescu and Usher, 2013). This occurred because larger inputs (whether congruent or incongruent) drove faster reaction times.

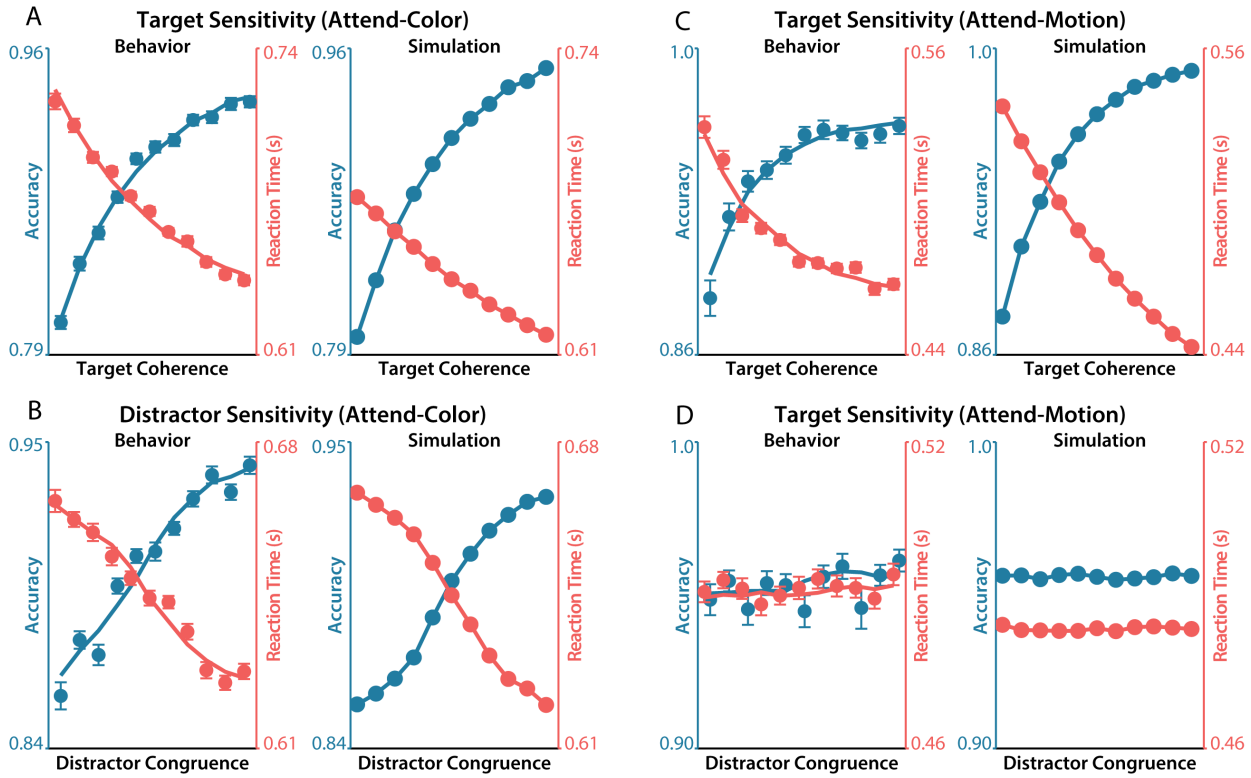


Figure 9. *Simulation of target and distractor sensitivity (see Figure 2).* **A-B)** Sensitivity to target coherence (A) and distractor congruence (B) in behavior (left) and in the FFIc simulation (right) for Attend-Color blocks. **C-D)** Same as A-B, but for Attend-Motion blocks.

We next used this model to test potential mechanisms underlying participants' within- and between-trial control adaptations. First, we tested whether participants' apparent within-trial dynamics in feature sensitivity plausibly resulted from actual within-trial changes in control gains governing feature sensitivity, or whether such dynamics could result from static control gains. We implemented time-varying feature gains as attractors with an initial gain (e.g., reflecting bottom-up salience or learning) that exponentially approaches a fixed point (e.g., determined by the task goals and control; cf. (Musslick et al., 2019; Steyvers et al., 2019)).

We found that incorporating these time-varying gains into our accumulator model allowed it to reproduce participants' behavioral dynamics. In accuracy, our model replicated the shift in target sensitivity over time, with the collapsing bound reducing performance on the slowest trials (Figure 10a). Our model similarly captured participants' decreased target sensitivity at later RTs (Figure 10b). Finally, our model recreated the analogous effects in RT, with faster errors for high coherence targets and incongruent distractors (Figure 10c-d). Critically, we were unable to replicate these qualitative patterns of behavior with FFI models in which control gains that were frozen throughout the trial (Supplementary Figure 4). Drift diffusion models with across-trial variability in gain, noise, or threshold; or drift diffusion models with within-trial dynamics in noise or threshold were similarly unable to capture our key effects without within-trial gain dynamics (Supplementary Figure 5).

At later RTs, participants were more likely to exhibit lapses in performance (i.e., choose randomly; ds = 0.25 to 0.93, $p < .001$, see Table 4), which were estimated with a separate term in our regression models (see 'Regression Analyses' section of Methods). This is evident in poorer overall performance in the slowest RT bin, relative to the 2^{nd}-4^{th} bins (see Figure 10a, left panel, pink line). A similar 'hook' is often observed in RT-conditioned accuracy functions, with gradually better performance followed by poorer performance for the slowest RTs (van den Wildenberg et al., 2010; Weichart et al., 2020). Our simulation captured this global reduction in accuracy by including a collapsing boundary (Drugowitsch et al., 2012; Rosenbaum et al., 2022), which leads to late errors irrespective of feature coherence (see Supplementary Figure 5 for contrast to fixed bound). Notably, even though overall accuracy is reduced over time, target sensitivity is stronger at the slowest RT bin relative to the earliest RT bin (compare navy and

pink psychometric slopes in Figure 10a), consistent with both feature-selective dynamics (gain control) and global dynamics (collapsing bound). By including a theory-driven mechanism for reductions in overall accuracy, our FFIc model captures performance trends in this slowest RT quantile that were difficult to capture with for more model-agnostic regression analyses.

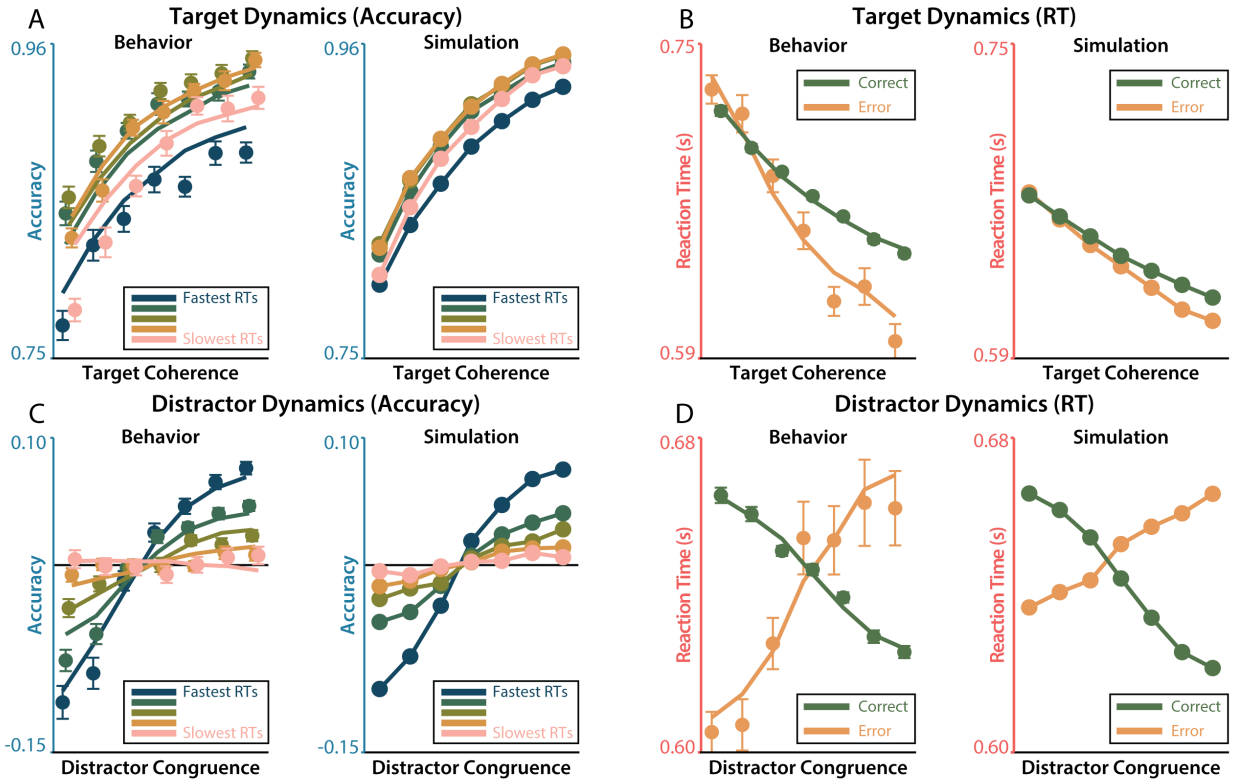


Figure 10. *Simulation of target and distractor sensitivity dynamics (see Figure 6).* **A-B)** RT-dependent (A) and accuracy-dependent (B) sensitivity to target coherence in behavior (left) and in the FFIc simulation (right). **C-D)** Same as A-B, but for distractor congruence.

The parallel feature pathways in this model are designed to capture the independent influences of a target and distractor information (Lindsay and Jacoby, 1994). However, the time-varying feature gains providing an account for the weak interactions we observed between target and

distractor sensitivity in accuracy. Despite there being no competition in feature processing in our model, we found these weak target-distractor interactions emerge in *simulated* accuracies but not simulated RTs. This interaction appeared to result from the different time courses of target and distractor sensitivity. As in participants' behavior, the model's errors due to incongruent distractors tend to occur early (Figure 10c-d), censoring target processing at a lower (early) level of sensitivity. This interplay between feature sensitivity *dynamics* (but not overall feature sensitivity per se) offers a plausible explanation for the subtle and seeming inconsistent interactions in participants' behavior.

Having provided an account of how each of our stimulus features is processed over the course of the trial depending on the task goal, we next tested a potential model-based account of the two forms of control adaptation we observed across trials. Our participants demonstrated enhanced target sensitivity on rewarded blocks, and suppressed distractor sensitivity after increasingly incongruent trials. In both cases, adaptation appeared to enhance sensitivity to stimulus features at the fastest reaction times.

To account for the early effects of conflict and incentives, we modified the initial conditions of our model's gain dynamics (Figure 11a). We simulated post-interference adaptation by initializing the distractor gain closer to its asymptote, and we simulated reward incentivization by initializing the target gain closer to its asymptote. We found that these simulations qualitatively reproduced participants' behavior, with stronger adaptation and reward effects earlier in the trial than later. The exponential dynamics in our attractor network parsimoniously accounts for the fact that dynamics tended to be faster when they were initialized further from the fixed point

(i.e., post-congruent trials). Thus, our model was able to capture the range of findings in this experiment: target-distractor sensitivity, within-trial dynamics, and how the dynamics of target and distractor processing may be influenced by control.

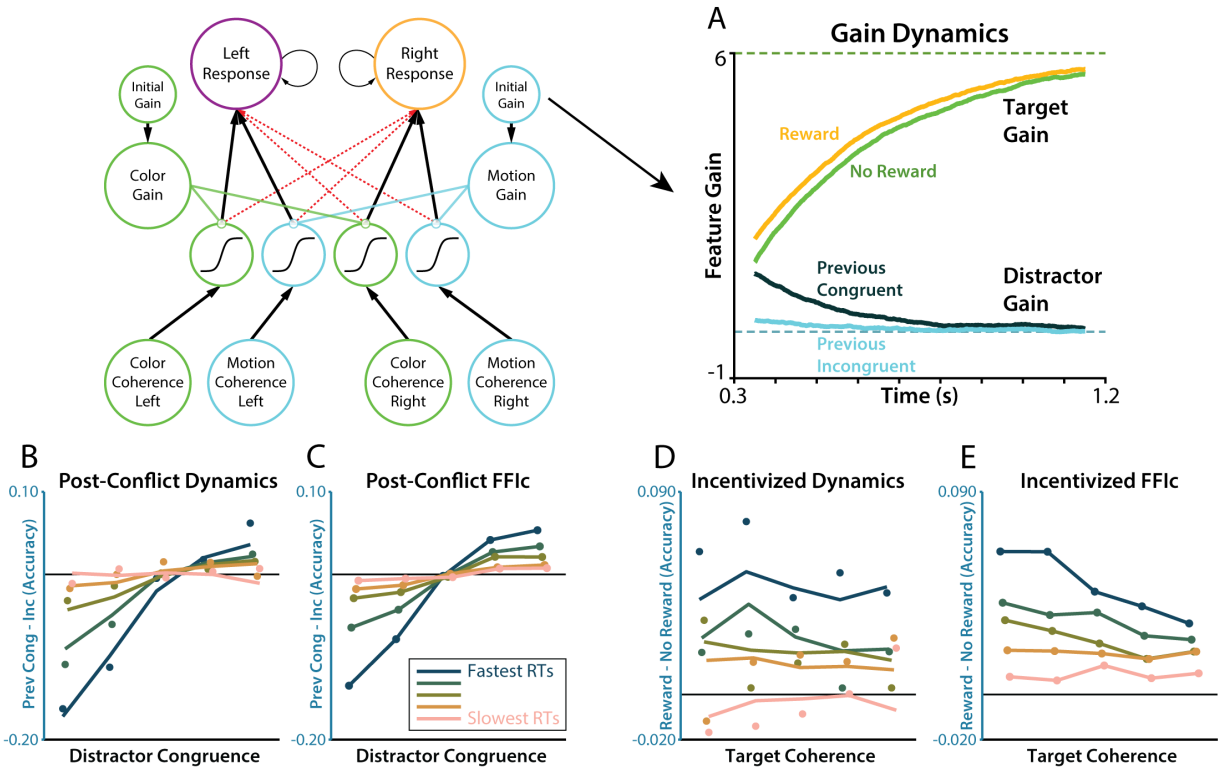


Figure 11. *Simulation of post-conflict and incentive effects (see Figure 7).* **A)** The influences of previous congruence (shade) and incentive effects (gold) were implemented through changes to the initial conditions of the feature gain dynamics, with previous congruence influencing initial distractor gain and incentives influence initial target gain. **B-C)** The influence of previous congruence on distractor sensitivity dynamics in behavior (B) and in the FFIc simulation (C). **D-E)** The influence of incentives on target sensitivity dynamics in behavior (D) and in the FFIc simulation €.

Discussion

When faced with distraction, we can sustain good performance by engaging with relevant information or ignoring disruptive information. Our experiment revealed that these strategies are under independent cognitive control, and are driven by distinct attentional dynamics. Using a bivalent random dot motion task with parametric target and distractor coherence (PACT), we found that target and distractor information have independent influences on participants' performance. Furthermore, we found that participants' sensitivity to targets and distractors was preferentially modulated by incentives and previous interference, respectively. These adaptations altered the initial conditions of feature-selective gains, which was followed by dynamic enhancement to target gains and suppression of distractor gains. These behavioral phenomena could be parsimoniously explained by a hybrid sequential sampling model with goal-dependent attractor dynamics over feature weights.

Together, these results support a cognitive control architecture that is parametric, multivariate, and dynamic. Previous research has found that cognitive effort is enhanced in response to incentives (Parro et al., 2018; Yee and Braver, 2018) and to previous conflict (Egner, 2007; Gratton et al., 1992). The current experiments extend these previous findings to show that these adaptations are both graded in their intensity, and selective in their allocation. These findings are consistent with a multivariate perspective on cognitive control (Ritz et al., 2022), in which people optimize a configuration of control signal according to their costs and benefits (Musslick et al., 2015; Shenhav et al., 2013). The target and distractor configurations observed here add to a body of work teasing apart the conditions under which people coordinate across multiple

control signals (Danielmeier et al., 2011; Leng et al., 2021; Noonan et al., 2016; Simen et al., 2009; Soutschek et al., 2015; Wöstmann et al., 2019).

A core question arising from these results is why there are preferential relationships between previous conflict with distractors, and incentives with targets. One possibility is that this is due to credit assignment. A system that could properly assign credit to features based on their contribution to conflict and incentives should allocate control towards distractors and targets. Distractors are a salient source of response conflict, and participants could adjust sensitivity to reduce this conflict. When participants were performing the more automatic Attend-Motion blocks, during which response conflict was absent, this adaptation was also absent. In contrast, reward contingencies were explicitly tied to target discrimination performance. During Attend-Motion blocks, there was a stronger association between target coherence and performance (e.g., due to response compatibility, and that only targets contributed to accuracy), potentially explaining why these blocks had larger incentive effects. This account is consistent with Bayesian models of cognitive control, such as those that predict feature congruence (Jiang et al., 2014; Yu et al., 2009) or the value of control policies (Bustamante et al., 2021; Lieder et al., 2018).

Our results also provide insight into the dynamic implementation of attentional control. Previous work has shown that within-trial attentional dynamics play an important role in both decision making (Callaway et al., 2021; Krajbich et al., 2010; Li and Ma, 2021; Westbrook et al., 2020) and cognitive control (Adkins and Lee, 2021; Hardwick et al., 2019; Servant et al., 2014; Ulrich et al., 2015; Weichart et al., 2020; White et al., 2011). These foundational experiments have

largely focused on spatial attention, with far less known about the dynamics of feature-based attention, where processing of targets and distractors is less mutually constrained. Whereas previous work has modeled within-trial dynamics as simplified impulse functions (Ulrich et al., 2015), our modeling approach extends these accounts with a more process-oriented focus on how a neural network could be parameterized to produce key patterns of within-trial attentional dynamics. Furthermore, relatively few experiments have studied how attentional dynamics are modified in response to control drivers like incentives or task demands (though see: (Adkins and Lee, 2021; van den Wildenberg et al., 2010; Yu et al., 2009)).

Our experiments show that the dynamics of target and distractor sensitivity are independent, and that previous conflict and incentives appear to operate through changes to the initial conditions of these feature gains[4]. These findings are broadly consistent with influential theories of attentional dynamics which propose that early task processing is largely driven by feature salience and statistical or reinforcement learning, whereas attentional control has a relatively slower timecourse ((Awh et al., 2012; Theeuwes, 2018, 2010), see also (van den Wildenberg et al., 2010))[5]. If participants are learning the relevance of different features, it's possible that these initial conditions in part reflect the prior probability that attention towards targets or distractors will support task goals (Lieder et al., 2018; Yu et al., 2009). Similar to how response priors are reflected in the initial decision state (Bogacz et al., 2006; Simen et al., 2009), priors on feature

[4] We found that just modifying feature gains' initial conditions parsimoniously accounted for incentive and previous conflict effects. Note that we do not explicitly compare this model to more complex models incorporating changes to parameters like decay rate and/or asymptotic gain, which should be more thoroughly investigated in future experiments.

[5] We assume that 'early' and 'late' processing do not reflect discrete stages (Hübner et al., 2010), but different timepoints in a continual process. While this is consistent with previous work showing that gradual attentional adjustments are a better model of flanker task performance (White et al., 2011), future work should experimentally confirm the continuous nature of these attentional dynamics.

priority may be reflected in the initial attentional state. In the case of previous interference, this could reflect learning whether distractors enhance performance (e.g., after trials on which congruent distractors led to better performance), or a local estimate of the probability a trial will be congruent (Yu et al., 2009). For incentives, this may reflect the expected target-reward contingency. Future research should investigate this account by measuring attentional dynamics as participants learn task contingencies (Shenhav et al., 2018).

Our patterns of conflict- and incentive-dependent dynamics help rule out stimulus-driven dynamics and support independent control over feature processing. After congruent trials, participants started the next trial with more similar target and distractor gains, that were then more quickly separated within the trial (Figure 7b). If these dynamics were an artifact of the decision process (e.g., due to accumulator attractors; (Wong and Wang, 2006), then we would expect that when target and distractor gains are initially more similar, there would be slower dynamics. Instead, we found faster dynamics, supporting a role for feedback control that reconfigures attentional gain to align with task goals. Additionally, during incentivized blocks, we saw that participants modified attentional dynamics for targets, but not distractors. This finding further supports the independence of these attentional dynamics, demonstrating that participants can alter attention towards individual features one at a time. This pattern of incentives enhancing sensitivity to target information, while also producing faster responding and a marginally higher lapse rate, is consistent with previous work on motivated attention. A recent experiment used drift diffusion modeling to show that participants increase their rate of evidence accumulation and decrease their response threshold when faced with higher rewards, consistent with the reward-rate optimal policy (Leng et al., 2021). The current experiment

extends these findings by revealing how specific attentional adjustments improve evidence accumulation, providing a more process-oriented account of motivated cognitive control.

Our dynamical process model may help link behavior in response conflict tasks to cognitive dynamics in other domains. In the domain of task-switching, recent cognitive models have developed similar dynamical accounts of how people reconfigure task sets. Classic work has shown that switch costs exponentially decay with preparation time (Monsell and Mizon, 2006; Rogers and Monsell, 1995), similar to the dynamics in the current task. Computational models have formalized these task set dynamics during the switch preparation period (Gilbert and Shallice, 2002; Jongkees et al., 2023; Musslick et al., 2019; Ueltzhöffer et al., 2015; Yeung and Monsell, 2003) and across trials (Grahek et al., 2022; Jaffe et al., 2023; Steyvers et al., 2019). If the within-trial dynamics we observe here reflect such "task set micro-adjustments" (Ridderinkhof, 2002), then our results highlight the computational similarities between different forms of cognitive flexibility. Both within trials and across tasks, reconfiguration appears to be well-captured by a common class of dynamical systems in which task configurations exponentially approach an appropriate set point. In this experiment, we show that these dynamics are multivariate and adjusted to meet local task demands through changes to initial conditions. Interestingly, control over initial conditions also plays a central role in the neural dynamics of motor preparation (Churchland et al., 2010; Kao et al., 2020; Remington et al., 2018), highlighting the broader similarities across motor and cognitive domains (Ritz et al., 2022, 2020) and generating neural predictions for the neural implementation of dynamic cognitive control.

The evidence we provide for dissociable control over target and distractor processing is consistent with previous neuroscience experiments that used neural correlates of stimulus

processing to argue for independent enhancement and suppression processes (Gazzaley et al., 2005; Noonan et al., 2016; Soutschek et al., 2015; Wöstmann et al., 2019). Our results extend these findings by exploring how different factors can contribute to dynamic reconfiguration of target and distractor attention, which we formalize in an explicit process model. Notably, our findings diverge from neuroimaging experiments that have suggested that control primarily acts through enhancements to target processing (Egner and Hirsch, 2005). One potential source of this divergence may be that people's control strategies differ depending on the source of task conflict (Braem et al., 2014; Egner, 2008; Egner et al., 2007). For example, tasks evoking stimulus-stimulus conflict (e.g., semantic competition in Stroop task) may require different strategies than tasks evoking stimulus-response conflict (e.g., distractors driving competing responses, as in PACT). Although previous work using Stroop-like tasks has found similar patterns of control adjustments as in the current experiment (Soutschek et al., 2015), this raises the broader question of whether the specific feature-control relationships in this experiment should generalize to other tasks. According to the Expected Value of Control theory, and the Learned Value of Control model that builds upon it, control strategies are adapted to specific task contexts (Lieder et al., 2018; Ritz et al., 2022; Shenhav et al., 2013). This framework predicts there will be strategic or learned control-feature mappings, rather than a rigid relationship between task features and control policies. The current results show that participants can independently control target and distractor processing when these features are independent, and future work should explore whether control strategies appropriately accommodate other tasks.

Interestingly, participants appeared to suppress distractor sensitivity even on congruent trials, evident in the right half of Figure 6d (see also (Mante et al., 2013; Pagan et al., 2022)),

suggesting that they are not reactively adjusting this control policy when the trial conditions deem it unnecessary or even detrimental. On its face, this finding presents a challenge to models that propose control allocation on the basis of response conflict (Botvinick et al., 2001; Yu et al., 2009), though much of the evidence for these theories comes from across-trial adjustments (Botvinick et al., 2001; Egner and Hirsch, 2005; Kerns et al., 2004; Yeung et al., 2004). The current results may thus inform understanding of the timescale over which people plan reactive control adjustments. In some cases, this decision process may take more time than would be helpful for fast within-trial reconfiguration.

Our analyses of attentional dynamics depend on participants' own response times and choices, raising concerns about selection biases (i.e., lack of experimental control over reaction times). While evidence accumulation modelling typically depends on choice-conditioned reaction times, inferring time-varying influence of targets and distractors presents a particular challenge. To address these concerns, we used simulations to show that the dynamic profiles we observed cannot be accounted for by an evidence accumulation model with static gains on target and distractor processing (Supplementary Figure 4) or models with dynamic changes to non-selective components like decision threshold (Supplementary Figure 5). Introducing dynamic feature gains allowed us to account for those same patterns (Figures 9 to 11; Supplementary Figure 5). These results are consistent with previous work validating DDM estimates of attentional dynamics in conflict tasks (White et al., 2018, 2011). Even if these measurements are valid, using sparse behavioral measures is an inefficient method for measuring latent dynamics, and may combine multiple processes (e.g., accumulation and threshold adjustments). By integrating across multiple convergent measures of decision and attentional dynamics – including interrogation protocols

(Adkins and Lee, 2021; Hardwick et al., 2019), motor tracking (Erb et al., 2016; Menceloglu et al., 2021; Scherbaum et al., 2010), and/or temporally-resolved neuroimaging (Fischer et al., 2018; Scherbaum et al., 2011; Weichart et al., 2020; Yeung et al., 2004) – future work can help strengthen and build on our understanding of continuous changes in the configuration of multiple control processes.

The evidence accumulation modeling in the current experiment was able to categorically rule out several alternative architectures, demonstrating the necessity and sufficiency of feature-specific adjustments for capturing the full array of putative attentional dynamics. Our model validation approach supports our interpretation of feature-selective adjustments, while committing less strongly to the specific formulation of attentional control (e.g., a specific model parameterization, or the functional form of the collapsing bound). An important direction for future research should be to leverage emerging methods for parameter estimation to directly fit our accumulator model to participants' behavior (Fengler et al., 2021; Weichart et al., 2020). This approach will help extend insights from the current experiment, such as enabling participant-specific parameters to reveal individual differences in attentional control.

Together, these experiments provide new insight into how we flexibly adapt to the changing demands of our environment. We find evidence for flexible control that aligns multiple forms of information processing with task goals, and can be captured by an computationally explicit process model. The developments from this experiment can help extend models of cognitive control towards richer accounts of how multivariate control configurations, such as across targets and distractors, are optimized during goal-directed behavior.

Supplementary Tables

Supplementary Table 1. Target and distractor sensitivity (Attend-Motion)					
DV	**Predictors**	**Exp 1 (df = 45) Effect size (*d*)**	**Exp 2 (df = 25) Effect size (*d*)**	**Exp 3 (df = 45) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Target coherence		4.39	5.44	**5.69×10^{-53}**
	Distractor congruence	0.732	0.352	0.269	**0.000916**
	Target × Distractor		0.0522	0.145	0.6210
RT	Target coherence		-1.47	-1.63	**2.71×10^{-21}**
	Distractor congruence	0.125	0.179	0.407	0.121
	Target × Distractor		0.0710	0.0371	0.864

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, α = 0.05) shown in bold.

Supplementary Table 2. Target and distractor sensitivity (Attend-Color - Attend-Motion)					
DV	**Predictors**	**Exp 1 Effect size (*d*)**	**Exp 2 Effect size (*d*)**	**Exp 3 Effect size (*d*)**	**Aggregate *p*-value**
Choice	Target coherence		-0.588 (df = 50.0)	-0.615 (df = 89.8)	**1.00×10^{-10}**
	Distractor congruence	0.432 (df = 88.9)	0.743 (df = 49.1)	0.971 (df = 75.8)	**5.79×10^{-19}**
RT	Target coherence		0.00957 (df = 44.4)	0.339 (df = 69.8)	0.182
	Distractor congruence	-0.804 (df = 78.8)	-1.19 (df = 42.8)	-1.38 (df = 59.1)	**2.69×10^{-31}**

Effect sizes are calculated from Welsh's contrasts across regression models. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, α = 0.05) shown in bold.

Supplementary Table 3. Correlations between RT and accuracy betas			
Experiment	**Correlands**	**MAP r-stat**	***p*-value**
Exp 1 (df = 54)	Distractor Betas	-0.891	**1.95×10^{-20}**

Exp 2 (df = 38)	Target Betas	-0.712	**1.27×10^{-7}**
	Distractor Betas	-0.875	**7.48×10^{-14}**
Exp 3 (df = 58)	Target Betas	-0.540	**4.20×10^{-6}**
	Distractor Betas	-0.907	**1.05×10^{-23}**

Parameter correlations are calculated from the MAP group-level parameter covariance. Statistically significant *p*-values (two-tailed, α = 0.05) are shown in bold.

Supplementary Table 4. Effects of previous conflict on feature sensitivity (Attend-Motion)

DV	**Predictors**	**Exp 1 (df = 41) Effect size (*d*)**	**Exp 2 (df = 19) Effect size (*d*)**	**Exp 3 (df = 39) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Distractor × Prev Distract	0.0113	-0.176	0.293	0.578
	target × Prev Target		-0.00795	-0.424	0.268
RT	Distractor × Prev Distract	0.503	-0.427	0.181	0.291
	target × Prev Target		0.131	-0.0333	0.777

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, α = 0.05) shown in bold.

Supplementary Table 5. Effects of previous conflict on feature sensitivity (Attend-Color - Attend-Motion)

DV	**Predictors**	**Exp 1 Effect size (*d*)**	**Exp 2 Effect size (*d*)**	**Exp 3 Effect size (*d*)**	**Aggregate *p*-value**
Choice	Distractor-dependent (Distractor - Distractor)	0.505 (df = 52.7)	0.984 (df = 33.6)	0.280 (df = 47.7)	**4.39×10^{-8}**
	Motion-dependent (Distractor - Target)		0.448 (df = 21.9)	0.651 (df = 40.2)	**2.27×10^{-5}**
RT	Distractor-dependent (Distractor - distractor)	-0.629 (df = 80.2)	-0.956 (df = 29.3)	-1.20 (df = 70.7)	**3.40×10^{-23}**
	Motion-dependent (Distractor - Target)		-0.237 (df = 19.1)	-0.0929 (df = 39.3)	0.337

Effect sizes are calculated from Welsh's contrasts across regression models. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, α = 0.05) shown in bold.

Supplementary Table 6. Effects of incentives on feature sensitivity (Attend-Motion)			
DV	**Predictors**	**Exp 3 (df = 41) Effect size (*d*)**	***p*-value**
Choice	Target × Reward	0.703	**6.02×10^{-5}**
	Distractor × Reward	0.289	0.110
Lapse Rate	Reward	-0.0696	0.670
RT	Target × Reward	-0.126	0.415
	Distractor × Reward	-0.192	0.265
	Reward	-0.0955	0.516

Effect sizes are calculated from MAP group-level regression estimates. Statistically significant *p*-values (two-tailed, $\alpha = 0.05$) are shown in bold.

Supplementary Table 7. Effects of incentives on feature sensitivity (Attend-Color – Attend-Motion)			
DV	**Predictors**	**Exp 3 Effect size (*d*)**	***p*-value**
Choice	Target × Reward	-0.279 (df = 59.0)	**0.0363**
	Distractor × Reward	-0.230 (df = 48.3)	0.117
RT	Target × Reward	0.0566 (df = 58.3)	0.667
	Distractor × Reward	0.271 (df = 77.0)	**0.0199**

Effect sizes are calculated from Welsh's contrasts across regression models. Statistically significant *p*-values (two-tailed, $\alpha = 0.05$) are shown in bold.

Supplementary Table 8. Dynamics of feature sensitivity across response times (Attend-Motion)					
DV	**Predictors**	**Exp 1 (df = 38) Effect size (*d*)**	**Exp 2 (df = 21) Effect size (*d*)**	**Exp 3 (df = 41) Effect size (*d*)**	**Aggregate *p*-value**
Choice	Target × RT		2.03	1.31	**3.30×10^{-20}**
	Distractor × RT	-0.257	0.194	-0.490	**0.0226**

Lapse Rate	RT	0.509	0.763	-0.234	0.941
RT	Target × Accuracy		0.772	0.679	**1.59×10^{-7}**
	Distractor × Accuracy	0.0758	-0.211	0.0199	0.931

Effect sizes are calculated from MAP group-level regression estimates. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

Supplementary Table 9. Dynamics of feature sensitivity across response times (Attend-Color – Attend-Motion)

DV	Predictors	Exp 1 Effect size (*d*)	Exp 2 Effect size (*d*)	Exp 3 Effect size (*d*)	Aggregate *p*-value
Choice	Target × RT		-2.08 (df = 23.0)	-1.09 (df = 46.1)	**3.35×10^{-17}**
	Distractor × RT	-0.102 (df = 56.3)	-0.585 (df = 22.4)	-0.539 (df = 78.5)	**4.57×10^{-5}**
RT	Target × Accuracy		-0.429 (df = 23.0)	-0.422 (df = 46.2)	**0.00148**
	Distractor × Accuracy	-0.344 (df = 66.4)	-0.648 (df = 29.6)	-0.723 (df = 78.6)	**4.56×10^{-11}**

Effect sizes are calculated from Welsh's contrasts across regression models. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

Supplementary Table 10. Model Collinearity

Model	Experiment	Accuracy Model Collinearity median [25% - 75%]	RT Model Collinearity median [25% - 75%]
Baseline	Experiment 1	1.4 [1.4 – 1.5]	1.1 [1.1 – 1.1]
	Experiment 2	1.4 [1.4 – 1.5]	1.1 [1.1 – 1.1]
	Experiment 3	1.4 [1.4 – 1.5]	1.1 [1.1 – 1.1]
Post-Conflict	Experiment 1	1.5 [1.4 – 1.5]	1.2 [1.1 – 1.3]
	Experiment 2	1.4 [1.4 – 1.5]	1.3 [1.2 – 1.3]
	Experiment 3	1.5 [1.4 – 1.5]	1.3 [1.2 – 1.3]
Reward	Experiment 3	1.4 [1.4 – 1.5]	1.2 [1.1 – 1.2]

Dynamics	Experiment 1	1.5 [1.5 – 1.6]	1.5 [1.3 – 2.0]
	Experiment 2	1.5 [1.4 – 1.5]	1.6 [1.4 – 1.7]
	Experiment 3	1.5 [1.4 – 1.5]	1.7 [1.5 – 2.0]
Post-Conflict Dynamics	Experiment 1	1.6 [1.6 – 1.8]	2.2 [1.7 – 3.3]
	Experiment 2	1.5 [1.4 – 1.5]	2.1 [1.8 – 2.4]
	Experiment 3	1.5 [1.5 – 1.6]	2.1 [1.7 – 2.5]
Reward Dynamics	Experiment 3	1.5 [1.5 – 1.6]	2.0 [1.7 – 2.4]

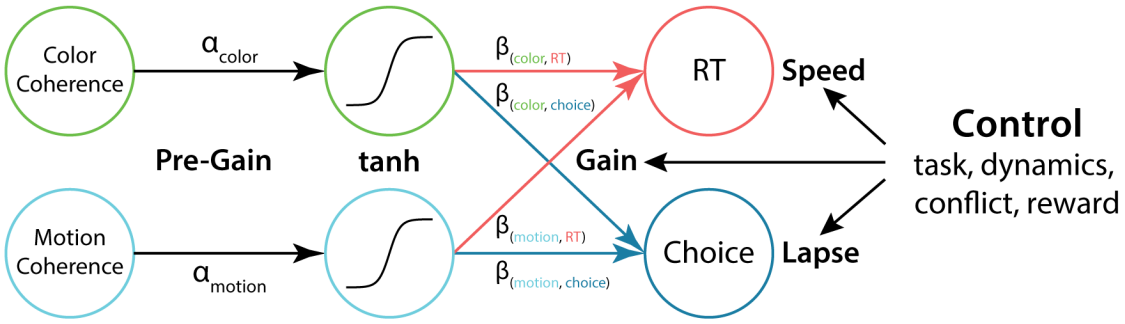
Belsley collinearity diagnostics for core models (from MATLAB's collintest). Diagnostic values are the ratio of the design matrix's largest singular value to its smallest singular value, summarized at different participant quantiles (i.e., median is the participant at the 50th percentile). A value of 1 is perfect orthogonality, and values below 30 are within the default tolerance. All values are well below 30, indicating tolerable collinearity.

Supplementary Table 11. Across-block changes in Feature Sensitivity Dynamics

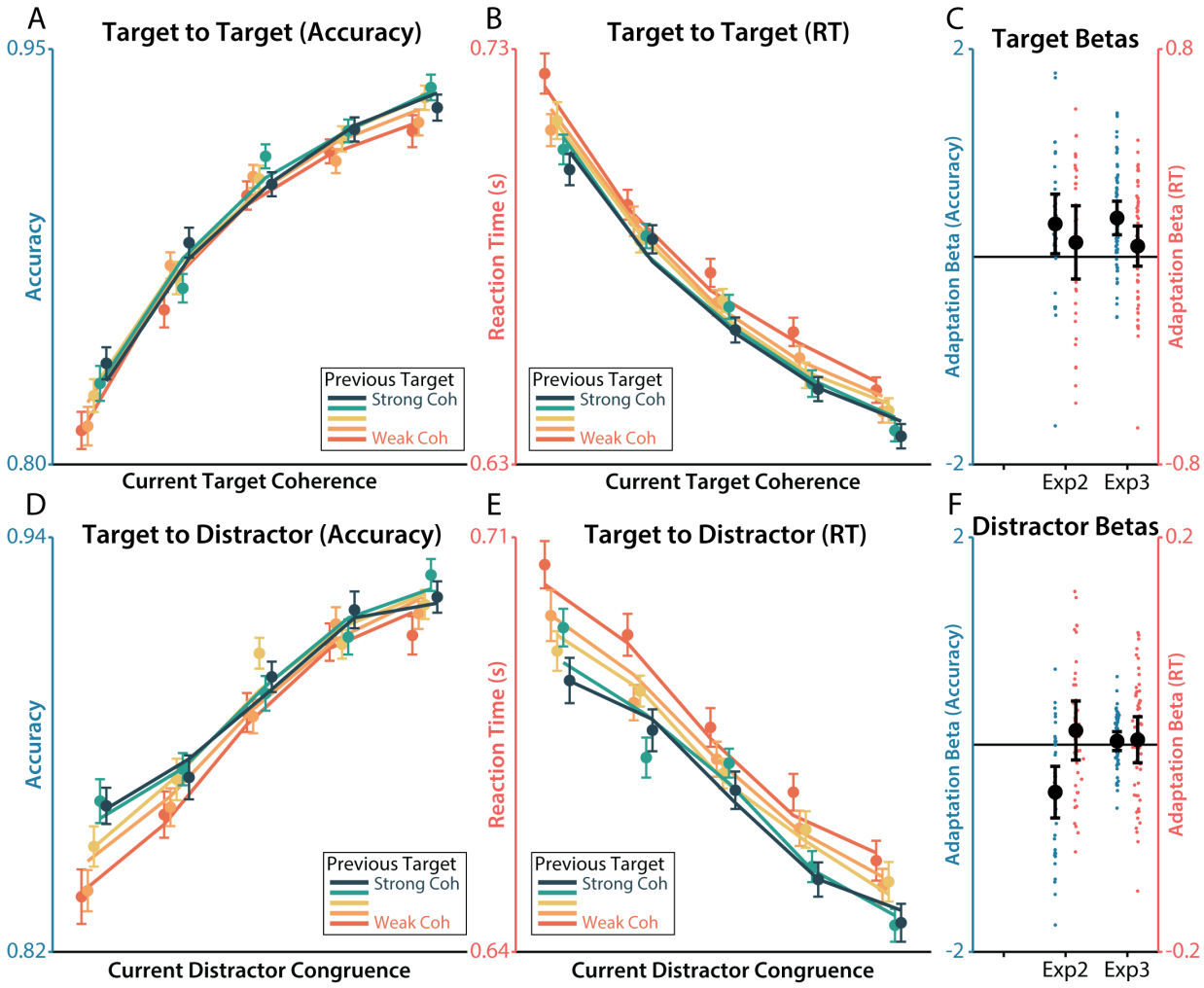
DV	**Predictors**	**Exp 2 Effect size (*d*)**	**Exp 3 Effect size (*d*)**	**Aggregate *p*-value**
Choice	Block × Distractor × RT	-0.501	-0.644	**5.21×10^{-6}**
	Block × Target × RT	0.348	0.480	**.000576**
	Block × Distractor	0.633	0.524	**1.27×10^{-5}**
	Block × Target	0.0698	-0.0581	.501
Lapse Rate	Block	0.380	0.0793	.175
	Block × RT	-0.505	-0.277	**.00342**
RT	Block × Distractor × Accuracy	0.168	-0.397	**.0399**
	Block × Target × Accuracy	-0.142	-0.153	**.286**
	Block × Distractor	0.0498	-0.129	.215
	Block × Target	0.371	0.475	**.000472**
	Block × Accuracy	0.396	0.448	**.000422**
	Block	-0.970	-1.23	**1.33×10^{-12}**

Effect sizes are calculated from Welsh's contrasts across regression models. *P*-values are aggregated across experiments, with statistically significant *p*-values (two-tailed, $\alpha = 0.05$) shown in bold.

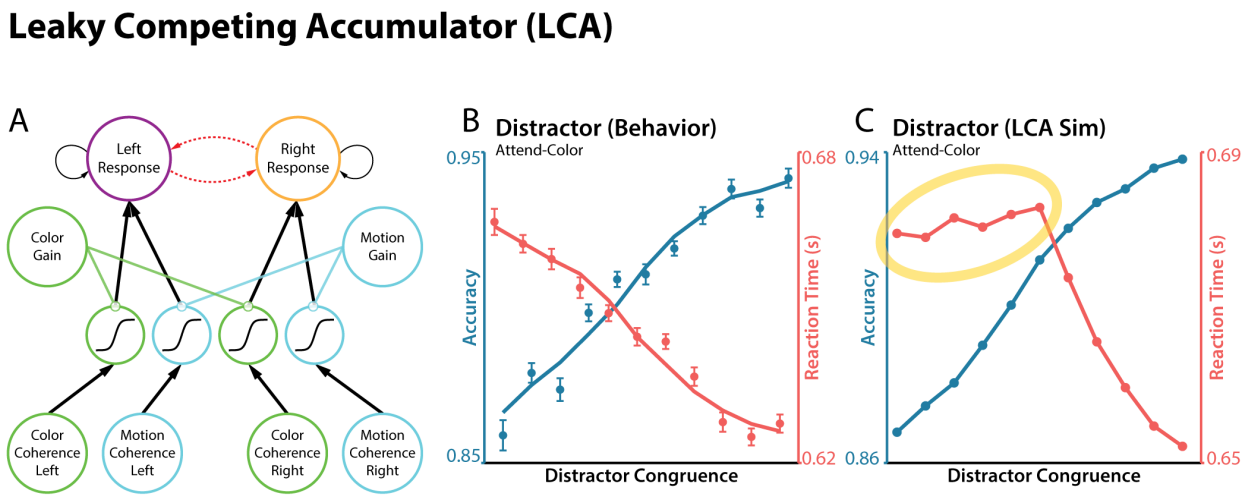
Supplementary Figures



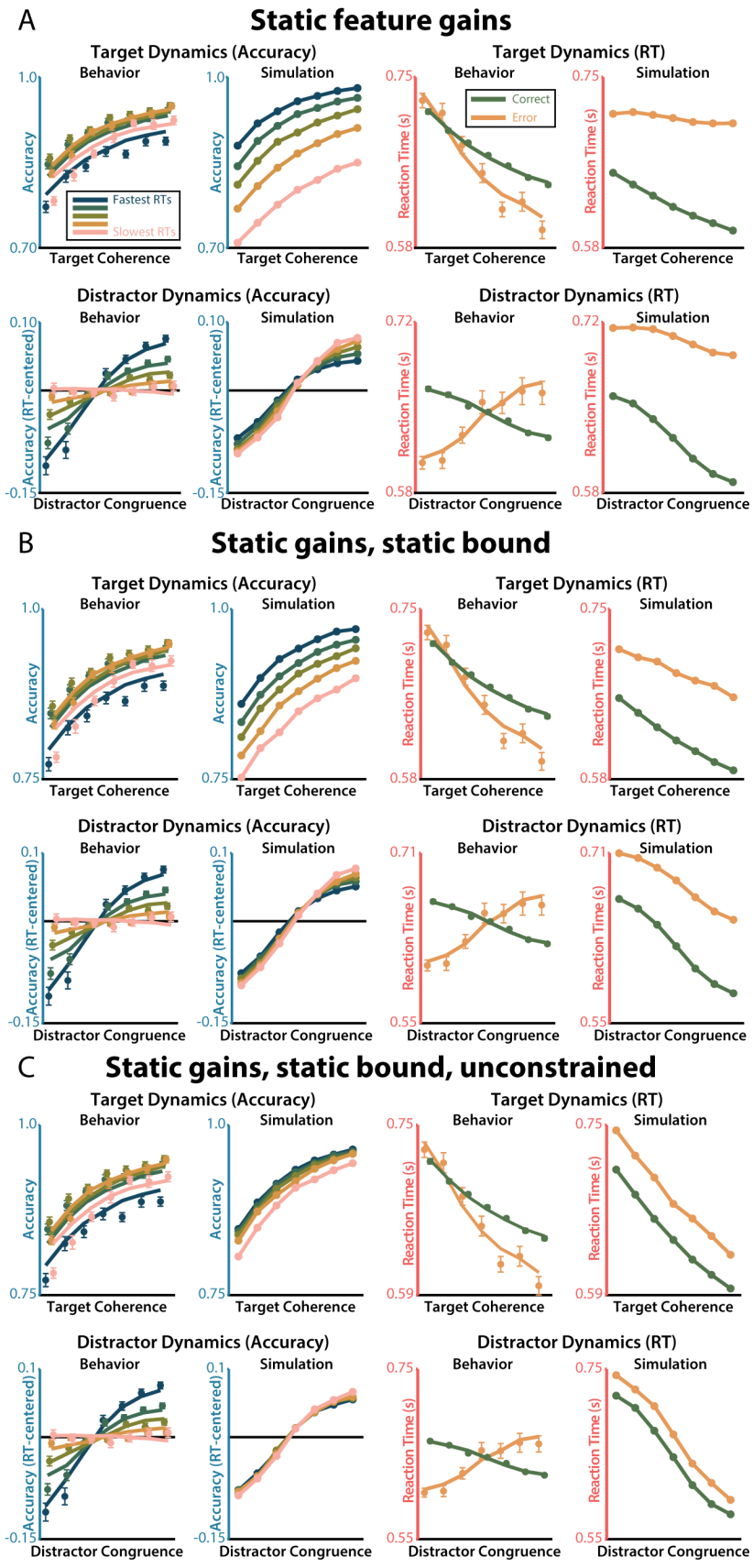
Supplementary Figure 1. *Regression schematic.* To estimate feature sensitivity, trial-specific color (green) and motion (cyan) coherence levels were passed through a hyperbolic tangent nonlinearity (tanh), with the α parameter determining the strength of the nonlinearity (see Methods). The linear relationships between transformed coherence and performance (RT in red and Choice in blue) were our estimates of participants' feature sensitivity. Our critical analyses tested whether potential indices of control (e.g., task instructions or incentives) moderated this feature sensitivity.



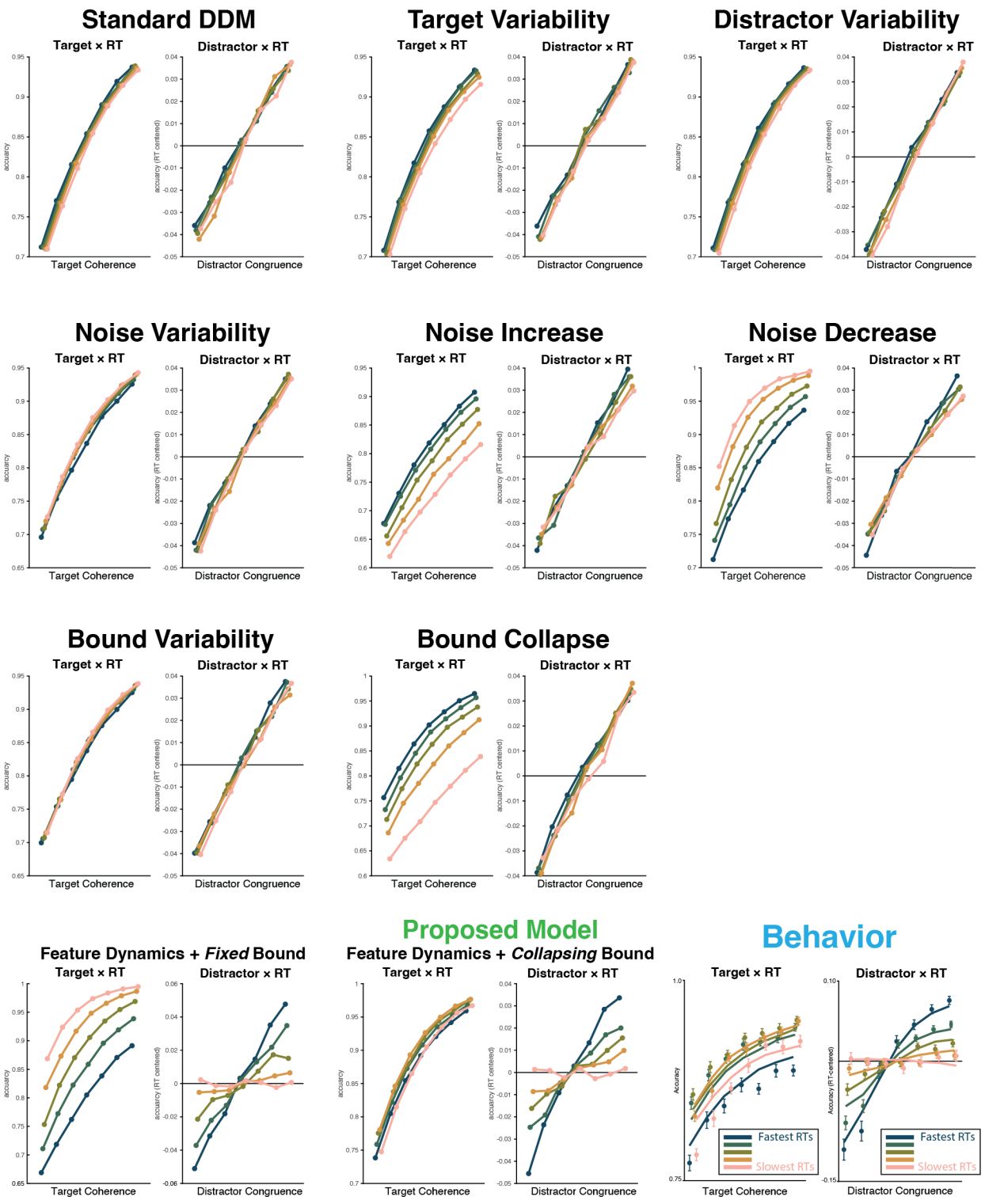
Supplementary Figure 2. *Target-dependent adaptation.* **A-B)** The relationship between target coherence and accuracy (A) was weaker when the previous trial had weaker target coherence (redder colors). There was not a significant effect for RT (B) Circles depict participant behavior and lines depict aggregated regression predictions. **C)** Regression estimates for the current target coherence by previous target coherence interaction, within each experiment. **D-E)** There was not a significant relationship between distractor congruence and previous target coherence in accuracy (D) or €(E). **F)** Regression estimates for the current distractor congruence by previous target coherence interaction, within each experiment. Error bars on behavior reflect within-participant SEM, error bars on regression coefficients reflect 95% CI.



Supplementary Figure 3. *Leak competing accumulator simulation.* **A)** We simulated behavior from a leaky competing accumulator (Usher and McClelland, 2001). In this model, the response accumulators directly compete. In our parameter regime, leak and competition parameters produce race-like accumulation dynamics (Bogacz et al., 2006; Weichart et al., 2020). **B-C)** We found that this parameter regime was unable to capture the effect of distractor congruence on reaction time, as stronger inputs (congruent or incongruent) produce faster RTs in a race-like regime (Teodorescu and Usher, 2013). Other parameter regime, producing DDM-like dynamics, would replicate our main simulation results (Bogacz et al., 2006).

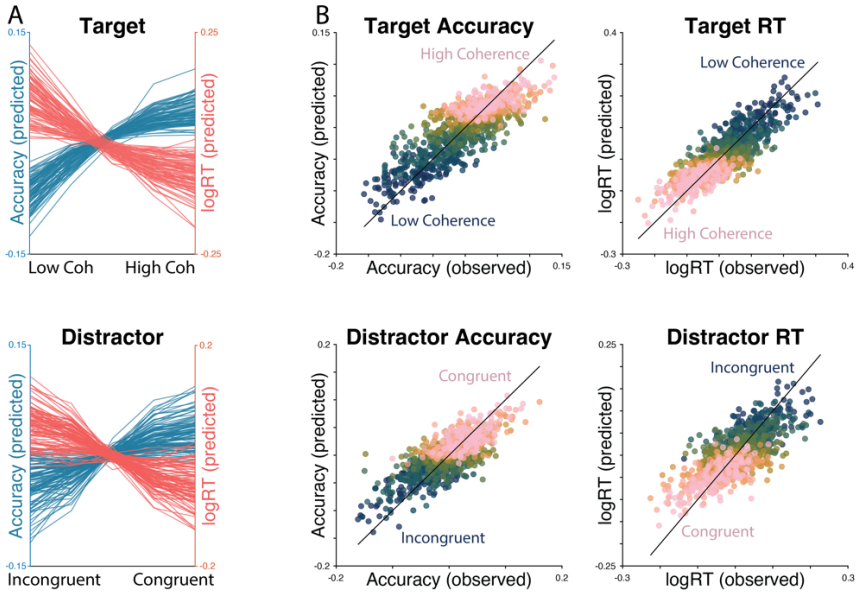

A
Static feature gains
Target Dynamics (Accuracy)
Behavior
Simulation
Accuracy
Fastest RTs
Slowest RTs
Target Coherence
Target Dynamics (RT)
Correct
Error
Reaction Time (s)
Distractor Dynamics (Accuracy)
Accuracy (RT-centered)
Distractor Congruence
Distractor Dynamics (RT)
B
Static gains, static bound
C
Static gains, static bound, unconstrained

Supplementary Figure 4. *Static feature gain simulations*. We simulated the FFI model under different formulations that lack feature sensitivity dynamics, showing that gain dynamics are necessary to capture the RT- and Accuracy-dependent feature sensitivity we observed in participants' behavior. Feature-specific processes are necessary to capture the opposite-going dynamics on target sensitivity and distractor sensitivity. A) Static model without feature dynamics. B) Static model without feature dynamics or collapse response threshold. C) Static model without feature dynamics, collapsing response threshold, or positive-rectified accumulators.

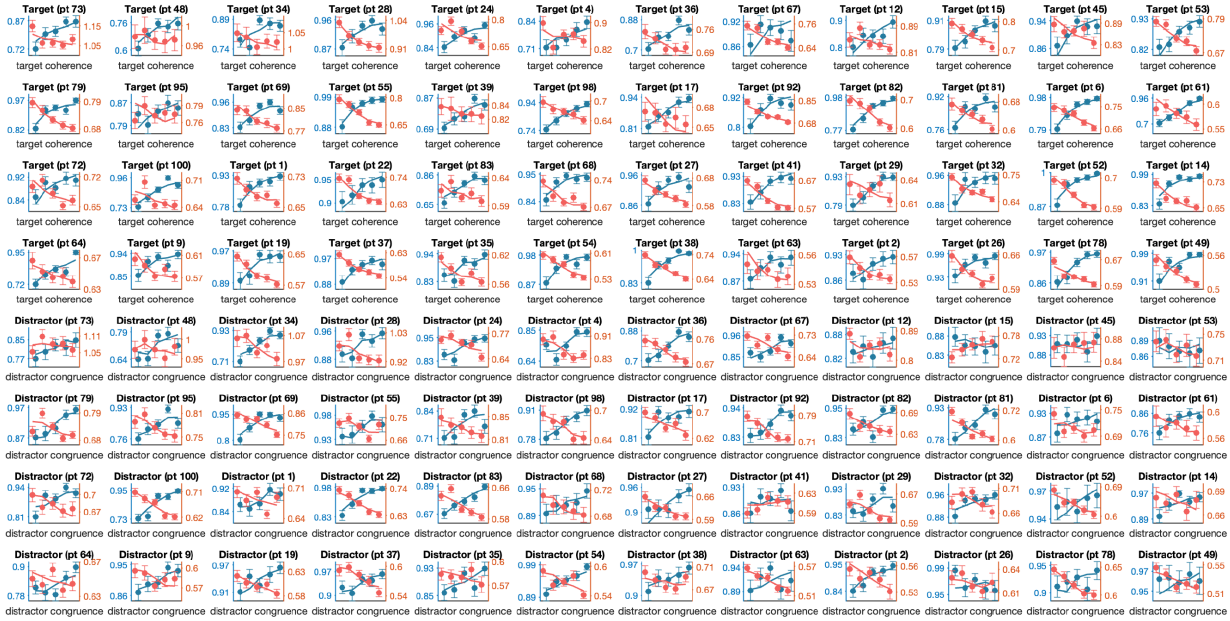


Supplementary Figure 5. *Dynamic drift diffusion simulations*. Drift diffusion model (DDM) simulations demonstrating the predictions from alternative formulations of within- and across-trial dynamics. Data are simulated target and distractor psychometric curves, conditioned on simulated RT quintiles (1 million simulations per

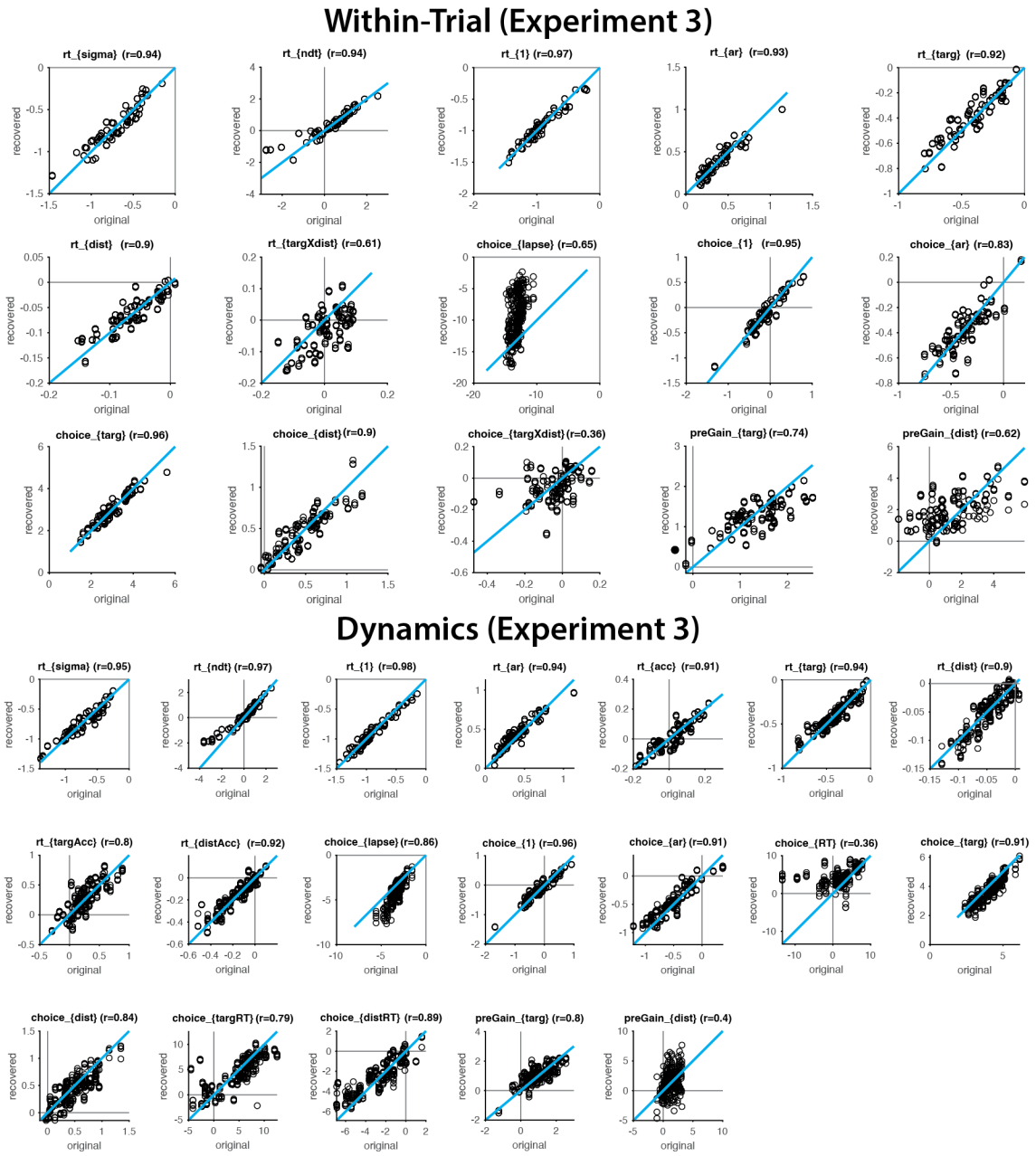
analysis). Row 1: Standard DDM, across-trial target gain variability, across-trial distractor gain variability. Row 2: across-trial accumulation noise variability, within-trial noise increase, within-trial noise decrease. Row 3: across-trial bound (threshold) variability, within-trial bound decrease ('collapsing bound'). Row 4: within-trial target gain enhancement and distractor suppression with fixed bound, within-trial target gain enhancement and distractor suppression with collapsing bound, participants' behavior. All simulations were performed using the dm package (package available at www.github.com/DrugowitschLab/dm; simulation scripts available at www.github.com/shenhavlab/PACT-public).



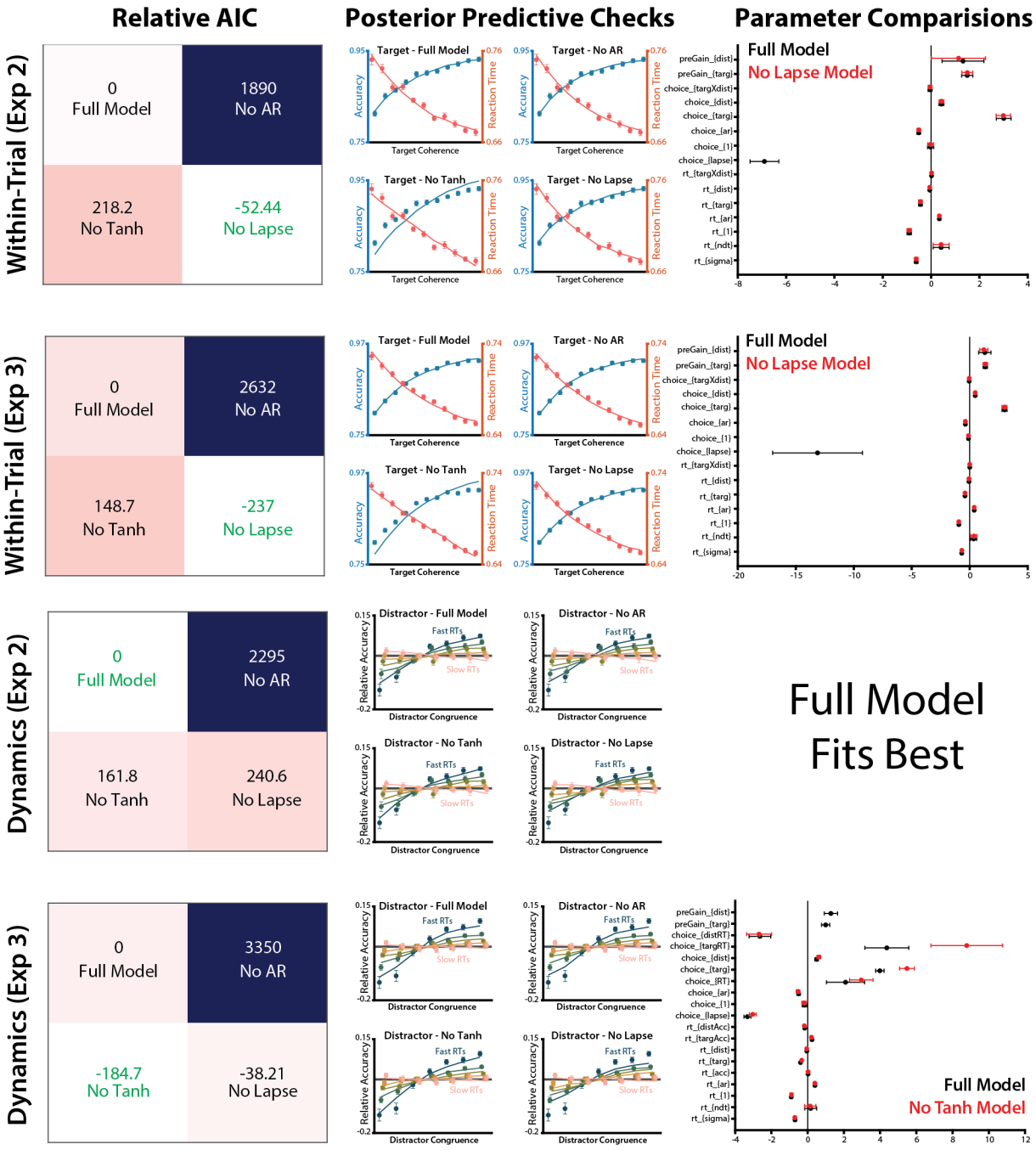
Supplementary Figure 6. *Aggregated posterior predictive checks.* **A)** Model predictions from participants in Experiments 2 and 3, showing predicted target sensitivity curves (top) and distractor sensitivity curves (bottom). Predictions are centered within-participant to remove individual intercepts. **B)** Model fit quality for participants in Experiments 2 and 3. Each participants' behavior (x-axis) is plotted against predicted behavior (y-axis), across five levels of target coherence (top) or distractor congruence (bottom; bluer to pinker indicates harder to easier conditions). Dots closer to the black identity reflect better model fit, and color gradients on y-axis reflect feature sensitivity. Predictions and behavior are centered within-participant to remove individual intercepts.



Supplementary Figure 7. *Single-participant posterior predictive checks.* Posterior predictive checks from 48 participants from Experiments 2 and 3, linearly spaced from the poorest model likelihood to the best model likelihood. First four rows are target sensitivity curves for accuracy (blue) and reaction time (red). Final four rows are distractor sensitivity curves (for the same participants) for accuracy (blue) and reaction time (red). Overlaid lines are single-trial model predictions aggregated like participants' behavior.

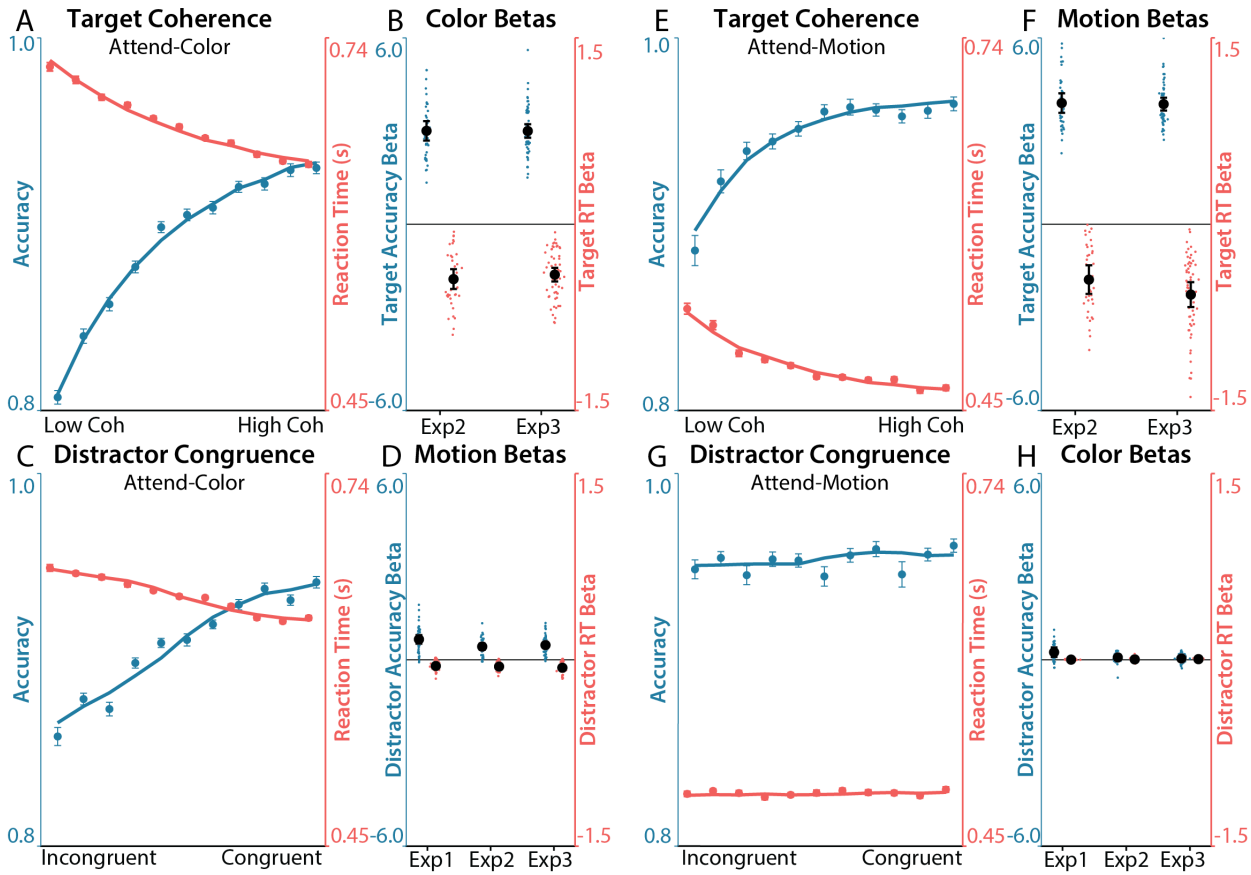


Supplementary Figure 8. *Parameter Recovery*. We simulated behavior from each participants' best-fitting parameters (x-axis) and then fit our model to this simulated behavior (y-axis). Each panel represents a parameter for the within-trial sensitivity model (top) and the within-trial dynamics model (bottom). Parameters were estimated hierarchically, with five simulated samples for each model (5 repetitions × 60 simulated participants). Gray horizontal and vertical lines reflect the parameter zero point, and the diagonal cyan line reflects the unity line. The simulated-recovered parameter correlation is reported in each panel title.



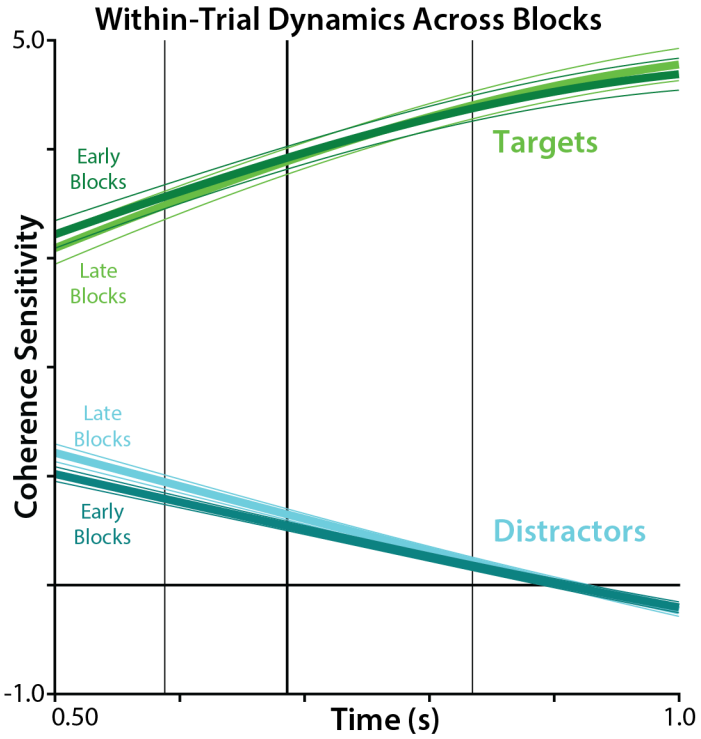
Supplementary Figure 9. *Parameter knock-out analysis*. Relative AIC (left column): parameter-penalized model fits for the regression model in the main text ('Full Model'), a model with previous RT and Choice removed ('No AR'), a model with tanh nonlinearities removed ('No Tanh'), and a model with the lapse rate response ('No Lapse'). Smaller values reflect better fit, with zero reflecting the AIC of the full model. Posterior predictive checks (center column): simulated behavior (lines) plotted over observed behavior (dots). Notice that removing tanh nonlinearities

fails to capture behavioral trends in within-subject models, and removing lapse terms fails to capture behavior in Dynamics models. Parameter comparisons (right column): model parameters plotted for the best-fitting model (red) and the full model (black). Notice that the parameters are very similar between these models, demonstrating that our key parameters are robust to knocking out other terms of the model.



Supplementary Figure 10. Target and distractor sensitivity (Equal Axes). A) Participants were more accurate (blue, left axis) and responded faster (red, right axis) when the target color had higher coherence. Lines depict aggregated regression predictions. In all graphs, behavior and regression predictions are averaged over participants and experiments. Data aggregated across Experiments 2 & 3. B) Regression estimates for the effect of target coherence on performance within each experiment, plotted for accuracy (blue, left axis) and RT (red, right axis). C) Participants were more accurate and responded faster when the distracting motion had higher congruence (coherence signed relative to target response). In all graphs, behavior and regression predictions are averaged over participants and experiments. Data aggregated across Experiments 1-3. D) Regression estimates for the effect of distractor

congruence on performance within each experiment, plotted for accuracy and RT. E-F) Similar to A-B, perfor€ce (E) and regression estimates (F) for the effects of target coherence during Attend-Motion blocks, in which motion was the target dimension. G-H) Similar to A-B, performance (G) and regression estimates (H) for the effects of distractor congruence during Attend-Motion blocks, in which color was the distractor dimension. Error bars on behavior reflect within-participant SEM, error bars on regression coefficients reflect 95% CI. Psychometric functions are jittered on the x-axis for ease of visualization. Y-axes have been equalized across features and tasks.



Supplementary Figure 11: Changes in within-trial dynamics across blocks. Compared to earlier blocks, in later blocks participants' earliest sensitivity was weaker for targets and stronger for distractors (i.e., less task-appropriate later in the experiment). However, participants also exhibited faster corrected dynamics in later blocks, showing similar sensitivity for the slowest reaction times.

Supplementary Note 1: Task Instructions

Motion training

You will see dots that are moving left or right. If the dots are moving left, respond with the left key. If the dots are moving right, respond with the right key. If you are correct, you will be told so, and if you make a mistake, you will be reminded about the response mappings. As always, please respond as quickly and accurately as you can.

Color training

You will see dots that are one of *these* four colors. If the dots are *these* colors, respond with *this* hand. If the dots are *these* colors, respond with *this* hand. If you are correct, you will be told so, and if you make a mistake, you will get to see the colors again. As always, please respond as quickly and accurately as you can.

Main Experiment

This is the main section. Now you will see dots that both have a color and are moving left or right. There will be two kinds of blocks. This block is a color block. In this block, you will have to respond to color with these keys, like you did in the training. You will no longer receive feedback. Other blocks will be motion blocks, and you will have to respond based on the direction of the dot motion. Feel free to take a short break between blocks and come get me after you've finished all the blocks. As always, please respond as quickly and accurately as you can. (Note: during experiments 2 and 3, we emphasized choosing the color that was in the majority).

Reward Variant

During some of the color and motion blocks, you will be able to earn a monetary reward based on your performance. This block is one of the HIGH reward blocks. These blocks will say ‘high reward’ at the

top, and the text will be gold. At the end of the experiment, we will randomly pick a bunch of trials from these blocks. Depending on how many trials that are fast and accurate, you will be able to earn up to $4. Other blocks will be 'NO reward' blocks, with 'NO reward' written at the top and white text. You will not earn any money for your performance on these blocks.

References

Adkins TJ, Lee T. 2021. Reward reduces habitual errors by enhancing the preparation of goal-directed actions. doi:10.31234/osf.io/hv9mz

Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM. 2018. Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *Am Psychol* **73**:3–25.

Awh E, Belopolsky AV, Theeuwes J. 2012. Top-down versus bottom-up attentional control: a failed theoretical dichotomy. *Trends Cogn Sci* **16**:437–443.

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* **113**:700–765.

Bogacz R, Usher M, Zhang J, McClelland JL. 2007. Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philos Trans R Soc Lond B Biol Sci* **362**:1655–1670.

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. *Psychol Rev* **108**:624.

Botvinick MM, Cohen JD. 2014. The computational and neural basis of cognitive control: charted territory and new frontiers. *Cogn Sci* **38**:1249–1285.

Braem S, Bugg JM, Schmidt JR, Crump MJC, Weissman DH, Notebaert W, Egner T. 2019. Measuring Adaptive Control in Conflict Tasks. *Trends Cogn Sci* **0**. doi:10.1016/j.tics.2019.07.002

Britten KH, Shadlen MN, Newsome WT, Movshon JA. 1992. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci* **12**:4745–4765.

Bustamante L, Lieder F, Musslick S, Shenhav A, Cohen J. 2021. Learning to Overexert Cognitive Control in a Stroop Task. *Cogn Affect Behav Neurosci*. doi:10.3758/s13415-020-00845-x

Callaway F, Rangel A, Griffiths TL. 2021. Fixation patterns in simple choice reflect optimal information sampling. *PLoS Comput Biol* **17**:e1008863.

Churchland MM, Cunningham JP, Kaufman MT, Ryu SI, Shenoy KV. 2010. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron* **68**:387–400.

Cohen JD, Dunbar K, McClelland JL. 1990. On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol Rev* **97**:332–361.

Cohen JD, Servan-Schreiber D, McClelland JL. 1992. A parallel distributed processing approach to automaticity. *Am J Psychol* **105**:239–269.

Danielmeier C, Eichele T, Forstmann BU, Tittgemeyer M, Ullsperger M. 2011. Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J Neurosci* **31**:1780–1789.

De Jong R, Liang CC, Lauber E. 1994. Conditional and unconditional automaticity: a dual-process model of effects of spatial stimulus-response correspondence. *J Exp Psychol Hum Percept Perform* **20**:731–750.

Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. 2012. The cost of accumulating evidence in perceptual decision making. *J Neurosci* **32**:3612–3628.

Egner T. 2008. Multiple conflict-driven control mechanisms in the human brain. *Trends Cogn Sci* **12**:374–380.

Egner T. 2007. Congruency sequence effects and cognitive control. *Cogn Affect Behav Neurosci* **7**:380–390.

Egner T, Hirsch J. 2005. Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nat Neurosci* **8**:1784–1790.

Erb CD, Moher J, Sobel DM, Song J-H. 2016. Reach tracking reveals dissociable processes underlying cognitive control. *Cognition* **152**:114–126.

Fischer AG, Nigbur R, Klein TA, Danielmeier C, Ullsperger M. 2018. Cortical beta power reflects decision dynamics and uncovers multiple facets of post-error adaptation. *Nat Commun* **9**:5038.

Giesen C, Frings C, Rothermund K. 2012. Differences in the strength of distractor inhibition do not affect distractor-response bindings. *Mem Cognit* **40**:373–387.

Gilbert SJ, Shallice T. 2002. Task switching: a PDP model. *Cogn Psychol* **44**:297–337.

Gold JI, Shadlen MN. 2007. The neural basis of decision making. *Annu Rev Neurosci* **30**:535–574.

Grahek I, Leng X, Fahey MP, Yee D, Shenhav A. 2022. Empirical and Computational Evidence for Reconfiguration Costs During Within-Task Adjustments in Cognitive Control. *Proceedings of the Annual Meeting of the Cognitive Science Society* **44**.

Gratton G, Coles MGH, Donchin E. 1992. Optimizing the use of information: strategic control of activation of responses. *J Exp Psychol Gen* **121**:480.

Hardwick RM, Forrence AD, Krakauer JW, Haith AM. 2019. Time-dependent competition between goal-directed and habitual response preparation. *Nat Hum Behav* **3**:1252–1262.

Hommel B, Proctor RW, Vu K-PL. 2004. A feature-integration account of sequential effects in the Simon task. *Psychol Res* **68**:1–17.

Hübner R, Steinhauser M, Lehle C. 2010. A dual-stage two-phase model of selective attention. *Psychol Rev* **117**:759–784.

Jaffe PI, Poldrack RA, Schafer RJ, Bissett PG. 2023. Modelling human behaviour in cognitive tasks with latent dynamical systems. *Nature Human Behaviour* 1–15.

Jiang J, Beck J, Heller K, Egner T. 2015. An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nat Commun* **6**:8165.

Jiang J, Heller K, Egner T. 2014. Bayesian modeling of flexible cognitive control. *Neurosci Biobehav Rev* **46 Pt 1**:30–43.

Jongkees B, Todd M, Lloyd K, Dayan P, Cohen JD. 2023. When it pays to be quick: dissociating control over task preparation and speed-accuracy trade-off in task switching. doi:10.31234/osf.io/quhns

Kang YH, Löffler A, Jeurissen D, Zylberberg A, Wolpert DM, Shadlen MN. 2021. Multiple decisions about one object involve parallel sensory acquisition but time-multiplexed evidence incorporation. *Elife* **10**. doi:10.7554/eLife.63721

Kao T-C, Sadabadi MS, Hennequin G. 2020. Optimal anticipatory control as a theory of movement preparation: a thalamo-cortical circuit model. *bioRxiv*. doi:10.1101/2020.02.02.931246

Kayser AS, Erickson DT, Buchsbaum BR, D'Esposito M. 2010. Neural representations of relevant and irrelevant features in perceptual decision making. *J Neurosci* **30**:15778–15789.

Kerns JG, Cohen JD, MacDonald AW 3rd, Cho RY, Stenger VA, Carter CS. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science* **303**:1023–1026.

Krajbich I, Armel C, Rangel A. 2010. Visual fixations and the computation and comparison of value in simple choice. *Nat Neurosci* **13**:1292–1298.

Krueger LE. 1989. Reconciling Fechner and Stevens: Toward a unified psychophysical law. *Behav Brain Sci* **12**:251–267.

Laming D. 1979. Autocorrelation of choice-reaction times. *Acta Psychol* **43**:381–412.

Lau B, Glimcher PW. 2005. Dynamic response-by-response models of matching behavior in rhesus monkeys. *J Exp Anal Behav* **84**:555–579.

Leng X, Yee D, Ritz H, Shenhav A. 2021. Dissociable influences of reward and punishment on adaptive cognitive control. *PLoS Comput Biol* **17**:e1009737.

Li Z-W, Ma WJ. 2021. An uncertainty-based model of the effects of fixation on choice. *PLoS Comput Biol* **17**:e1009190.

Lieder F, Shenhav A, Musslick S, Griffiths TL. 2018. Rational metareasoning and the plasticity of cognitive control. *PLoS Comput Biol* **14**:e1006043.

Lindsay DS, Jacoby LL. 1994. Stroop process dissociations: the relationship between facilitation and interference. *J Exp Psychol Hum Percept Perform* **20**:219–234.

Lipták T. 1958. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*

3:171–197.

Mante V, Sussillo D, Shenoy KV, Newsome WT. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**:78–84.

Mayr U, Awh E, Laurey P. 2003. Conflict adaptation effects in the absence of executive control. *Nat Neurosci* **6**:450–452.

Menceloglu M, Suzuki S, Song J-H. 2021. Revealing the effects of temporal orienting of attention on response conflict using continuous movements. *Atten Percept Psychophys*. doi:10.3758/s13414-020-02235-4

Moeller B, Frings C. 2014. Attention meets binding: only attended distractors are used for the retrieval of event files. *Atten Percept Psychophys* **76**:959–978.

Monsell S, Mizon GA. 2006. Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? *J Exp Psychol Hum Percept Perform* **32**:493–516.

Musslick S, Bizyaeva A, Agaron S, Leonard N, Cohen JD. 2019. Stability-flexibility dilemma in cognitive control: a dynamical system perspectiveProceedings of the 41st Annual Meeting of the Cognitive Science Society.

Musslick S, Jang SJ, Shvartsman M, Shenhav A, Cohen JD. 2018. Constraints associated with cognitive control and the stability-flexibility dilemmaCogSci. shenhavlab.org.

Musslick S, Shenhav A, Botvinick M, Cohen J. 2015. A Computational Model of Control Allocation based on the Expected Value of Control2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making. Presented at the Multidisciplinary Conference on Reinforcement Learning and Decision Making.

Nieder A, Miller EK. 2003. Coding of cognitive magnitude: compressed scaling of numerical

information in the primate prefrontal cortex. *Neuron*.

Noonan MP, Adamian N, Pike A, Printzlau F, Crittenden BM, Stokes MG. 2016. Distinct Mechanisms for Distractor Suppression and Target Facilitation. *J Neurosci* **36**:1797–1807.

Norman DA, Bobrow DG. 1975. On data-limited and resource-limited processes. *Cogn Psychol* **7**:44–64.

Pagan M, Tang VD, Aoi MC, Pillow JW, Mante V, Sussillo D, Brody CD. 2022. A new theoretical framework jointly explains behavioral and neural variability across subjects performing flexible decision-making. *bioRxiv*. doi:10.1101/2022.11.28.518207

Parro C, Dixon ML, Christoff K. 2018. The neural basis of motivational influences on cognitive control. *Hum Brain Mapp* **39**:5097–5111.

Posner M, Snyder C. 1975. Attention and cognitive control.

Ratcliff R, McKoon G. 2008. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput* **20**:873–922.

Remington ED, Egger SW, Narain D, Wang J, Jazayeri M. 2018. A Dynamical Systems Perspective on Flexible Motor Timing. *Trends Cogn Sci* **22**:938–952.

Ridderinkhof KR. 2002. Micro- and macro-adjustments of task set: activation and suppression in conflict tasks. *Psychol Res* **66**:312–323.

Ritz H, Frömer R, Shenhav A. 2020. Bridging Motor and Cognitive Control: It's About Time! *Trends Cogn Sci*.

Ritz H, Leng X, Shenhav A. 2022. Cognitive Control as a Multivariate Optimization Problem. *J*

Cogn Neurosci **34**:569–591.

Rogers RD, Monsell S. 1995. Costs of a predictible switch between simple cognitive tasks. *J Exp Psychol Gen* **124**:207.

Rosenbaum D, Glickman M, Fleming SM, Usher M. 2022. The Cognition/Metacognition Trade-Off. *Psychol Sci* 9567976211043428.

Scherbaum S, Dshemuchadse M, Fischer R, Goschke T. 2010. How decisions evolve: the temporal dynamics of action selection. *Cognition* **115**:407–416.

Scherbaum S, Fischer R, Dshemuchadse M, Goschke T. 2011. The dynamics of cognitive control: evidence for within-trial conflict adaptation from frequency-tagged EEG. *Psychophysiology* **48**:591–600.

Schmidt JR. 2019. Evidence against conflict monitoring and adaptation: An updated review. *Psychon Bull Rev* **26**:753–771.

Schmidt JR, De Houwer J. 2011. Now you see it, now you don't: controlling for contingencies and stimulus repetitions eliminates the Gratton effect. *Acta Psychol* **138**:176–186.

Servant M, Montagnini A, Burle B. 2014. Conflict tasks and the diffusion framework: Insight in model constraints based on psychological laws. *Cogn Psychol* **72**:162–195.

Shadlen MN, Newsome WT. 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J Neurophysiol* **86**:1916–1936.

Shenhav A, Botvinick MM, Cohen JD. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron* **79**:217–240.

Shenhav A, Straccia MA, Musslick S, Cohen JD, Botvinick MM. 2018. Dissociable neural

mechanisms track evidence accumulation for selection of attention versus action. *Nat Commun* **9**:2485.

Shiffrin RM, Schneider W. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychol Rev* **84**:127.

Simen P, Contreras D, Buck C, Hu P, Holmes P, Cohen JD. 2009. Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. *J Exp Psychol Hum Percept Perform* **35**:1865–1897.

Soutschek A, Stelzel C, Paschke L, Walter H, Schubert T. 2015. Dissociable effects of motivation and expectancy on conflict processing: an fMRI study. *J Cogn Neurosci* **27**:409–423.

Stafford T, Ingram L, Gurney KN. 2011. Piéron's Law holds during stroop conflict: insights into the architecture of decision making. *Cogn Sci* **35**:1553–1566.

Steyvers M, Hawkins GE, Karayanidis F, Brown SD. 2019. A large-scale analysis of task switching practice effects across the lifespan. *Proc Natl Acad Sci U S A*. doi:10.1073/pnas.1906788116

Stins JF, Polderman JCT, Boomsma DI, de Geus EJC. 2008. Conditional accuracy in response interference tasks: Evidence from the Eriksen flanker task and the spatial conflict task. *Adv Cogn Psychol* **3**:409–417.

Teodorescu AR, Usher M. 2013. Disentangling decision models: from independence to competition. *Psychol Rev* **120**:1–38.

Teufel HJ, Wehrhahn C. 2000. Evidence for the contribution of S cones to the detection of flicker brightness and red-green. *J Opt Soc Am A Opt Image Sci Vis* **17**:994–1006.

Theeuwes J. 2018. Visual Selection: Usually Fast and Automatic; Seldom Slow and Volitional. *J Cogn* **1**:29.

Theeuwes J. 2010. Top–down and bottom–up control of visual selection. *Acta Psychol* .

Tzelgov J, Henik A, Berger J. 1992. Controlling Stroop effects by manipulating expectations for color words. *Mem Cognit* **20**:727–735.

Ueltzhöffer K, Armbruster-Genç DJN, Fiebach CJ. 2015. Stochastic Dynamics Underlying Cognitive Stability and Flexibility. *PLoS Comput Biol* **11**:e1004331.

Ulrich R, Schröter H, Leuthold H, Birngruber T. 2015. Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions. *Cogn Psychol* **78**:148–174.

Urai AE, de Gee JW, Tsetsos K, Donner TH. 2019. Choice history biases subsequent evidence accumulation. *Elife* **8**. doi:10.7554/eLife.46331

Usher M, McClelland JL. 2001. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol Rev* **108**:550–592.

van den Wildenberg WPM, Wylie SA, Forstmann BU, Burle B, Hasbroucq T, Ridderinkhof KR. 2010. To head or to heed? Beyond the surface of selective action inhibition: a review. *Front Hum Neurosci* **4**:222.

Vogel TA, Savelson ZM, Otto AR, Roy M. 2020. Forced choices reveal a trade-off between cognitive effort and physical pain. *Elife* **9**. doi:10.7554/eLife.59410

Weichart ER, Turner BM, Sederberg PB. 2020. A model of dynamic, within-trial conflict resolution for decision making. *Psychol Rev*. doi:10.1037/rev0000191

Westbrook A, van den Bosch R, Määttä JI, Hofmans L, Papadopetraki D, Cools R, Frank MJ. 2020. Dopamine promotes cognitive effort by biasing the benefits versus costs of cognitive work. *Science* **367**:1362–1366.

White CN, Ratcliff R, Starns JJ. 2011. Diffusion models of the flanker task: discrete versus gradual attentional selection. *Cogn Psychol* **63**:210–238.

White CN, Servant M, Logan GD. 2018. Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychon Bull Rev* **25**:286–301.

Wichmann FA, Hill NJ. 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. *Percept Psychophys* **63**:1293–1313.

Wong K-F, Wang X-J. 2006. A recurrent network mechanism of time integration in perceptual decisions. *J Neurosci* **26**:1314–1328.

Wöstmann M, Alavash M, Obleser J. 2019. Alpha Oscillations in the Human Brain Implement Distractor Suppression Independent of Target Selection. *J Neurosci* **39**:9797–9805.

Wöstmann M, Störmer VS, Obleser J, Addleman DA, Andersen S, Gaspelin N, Geng J, Luck SJ, Noonan M, Slagter HA, al. E. 2021. Ten simple rules to study distractor suppression. doi:10.31234/osf.io/vu2k3

Wylie SA, Ridderinkhof KR, Elias WJ, Frysinger RC, Bashore TR, Downs KE, van Wouwe NC, van den Wildenberg WPM. 2010. Subthalamic nucleus stimulation influences expression and suppression of impulsive behaviour in Parkinson's disease. *Brain* **133**:3611–3624.

Yee DM, Braver TS. 2018. Interactions of Motivation and Cognitive Control. *Curr Opin Behav Sci* **19**:83–90.

Yeung N, Botvinick MM, Cohen JD. 2004. The Neural Basis of Error Detection: Conflict Monitoring and the Error-Related Negativity. *Psychol Rev* **111**:931–959.

Yeung N, Monsell S. 2003. Switching between tasks of unequal familiarity: the role of stimulus-attribute and response-set selection. *J Exp Psychol Hum Percept Perform* **29**:455–469.

Yu AJ, Dayan P, Cohen JD. 2009. Dynamics of attentional selection under conflict: toward a rational Bayesian account. *J Exp Psychol Hum Percept Perform* **35**:700–717.

Zaykin DV. 2011. Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *J Evol Biol* **24**:1836–1841.

Zhang J, Kornblum S. 1997. Distributional analysis and De Jong, Liang, and Lauber's (1994) dual-process model of the Simon effect. *J Exp Psychol Hum Percept Perform* **23**:1543–1551.